Evaluation of Cloud and Water Vapor Simulations in IPCC AR5 Climate Models Using NASA "A-Train" Satellite Observations

Jonathan H. Jiang, Hui Su, Chengxing Zhai, Vincent S. Perun
Jet Propulsion Laboratory (JPL), California Institute of Technology, Pasadena, California, USA

Anthony Del Genio and Larissa S. Nazarenko
Goddard Institute for Space Studies (GISS), New York, New York, USA

Leo J. Donner, Larry Horowitz, Charles Seman
Geophysical Fluid Dynamics Laboratory (GFDL), Princeton, New Jersey, USA

Jason Cole
Canadian Centre for Climate Modeling and Analysis (CCCMA), Environment Canada, Toronto, Canada

Andrew Gettelman
National Center for Atmospheric Research (NCAR), Boulder, Colorado

Mark Ringer
UK Met Office (UKMO) Hadley Center, London, UK

Leon Rotstayn
Commonwealth Scientific and Industrial Research Organisation (CSIRO), Campbell, Australia

Stephen Jeffrey
Queensland Climate Change Centre of Excellence (QCCCE), Queensland, Australia

Tongwen Wu
Beijing Climate Center (BCC), China Meteorological Administration, Beijing, China

Florent Brient and Jean-Louis Dufresne
Laboratory of Dynamical Meteorology, Institute Pierre Simon Laplace (IPSL), France

Hideaki Kawai and Tsuyoshi Koshiro
Meteorological Research Institute (MRI), Japan Meteorological Agency, Tsukuba, Japan

Masahiro Watanabe
Model for Interdisciplinary Research On Climate (MIROC)
Atmospheric and Ocean Research Institute, University of Tokyo, Chiba, Japan

Tristan LÉcuyer
University of Wisconsin-Madison, Madison, Wisconsin, USA

William G. Read, Joe W. Waters, Baijun Tian, Joao P. Teixeira, Graeme L. Stephens
JPL, California Institute of Technology, Pasadena, California, USA

Key Words: Clouds, Water Vapor, Climate Model, and Satellite Observation

Abstract

Using NASA's A-Train satellite measurements, we evaluate the accuracy of cloud water content (CWC) and water vapor mixing ratio (H_2O) outputs from 19 climate models submitted to the Intergovernmental Panel for Climate Change (IPCC) Fifth Assessment Report (AR5). We find improvements in 8 AR5 models for cloud water path relative to their counterparts for the IPCC Fourth Assessment Report. For vertical structures of CWC and H_2O, we find that the model spreads and their differences from the observations are larger in the upper troposphere (UT) than in the lower and mid-troposphere (LMT). The modeled tropical oceanic mean CWCs (H_2Os) range from ~3% to ~15× (~1% to 2×) of the observations in the UT and 40% to 2× (within 10%) of the observations in the LMT. The spatial distributions of clouds at 215 hPa are relatively well-correlated with observations, noticeably better than those for the LMT clouds. Although both water vapor and clouds are better simulated in the LMT than in the UT, there is no apparent correlation between the model biases in clouds and water vapor. Numerical scores are used to compare different AR5 model performances in regards to spatial mean, spatial variance and spatial distribution of CWC and H_2O at 100, 215, 600 and 900 hPa pressure levels. Model performances at each pressure level are ranked according a simple average of all scores for that pressure level, and overall performances are ranked according to a simple average of the scores for all four pressure levels.

1. Introduction

IPCC projections of climate change currently rely on some 20 climate models' simulations conducted at climate research centers worldwide. The outputs of these models consist of climate change indicators such as temperature, precipitation, clouds and water vapor. Clouds (both ice and liquid) and water vapor, which we consider here, are important modulators of climate and are involved in feedbacks that strongly affect global circulation and energy balance. Both ice and liquid clouds significantly affect the radiation budget through their shortwave albedo and longwave greenhouse effects [e.g. *Randall and Tjemkes,* 1991]. Water vapor produces the most important positive feedback affecting climate change [e.g. *Soden and Held,* 2006]. Small errors or uncertainties in water vapor simulations can cause large errors or uncertainties in predicting climate change - even though current models generally agree on the sign and magnitude of water vapor feedback [*Held and Soden*, 2000; *Soden and Held*, 2006]. Convective parameterizations, and their uncertainties, make difficult the accurate model simulation of water vapor and clouds. Uncertainties in couplings between clouds and water vapor also add uncertainty to climate change predictions. Cloud feedback remains the largest source of uncertainty in predicting climate change [e.g. *Cess et al*., 1996; *Soden and Held,* 2006; *Bony et al.,* 2006; *Randall et al.,* 2007; *Waliser et al.,* 2009]. Improving the accuracy of cloud and water vapor simulations by climate models is thus of critical importance.

Climate modelers have, over the past decade, undertaken tremendous efforts to improve model representation of clouds and water vapor, using a variety of observations to guide their work. ISCCP, ERBE, SSM/I, TRMM, NVAP and other satellite data for clouds and water vapor were used prior to 2002. The A-Train satellite constellation [*L'Ecuyer and Jiang*, 2010], which began in 2002, gives a significant improvement by providing co-located and near-

simultaneous vertical profiles of clouds and water vapor and presents the first 3-dimensional simultaneous global observations of these important parameters. The A-Train observations place, more than previously possible, stringent constraints on model simulations of clouds and water vapor, and can help identify specific model outputs that either are modeled adequately or need improvement.

We here - in the first of a series of papers to address the cloud and water vapor performance of climate models submitted to the Intergovernmental Panel for Climate Change (IPCC) Fifth Assessment Report (AR5) - compare multi-year means from A-Train observations with those from the AR5, and previous IPCC AR4, models. Global and zonal (tropical, mid-latitude, and high latitude) multi-year spatial means and spatial distributions are considered. Attention is given to vertical structure and the combined evaluation of cloud and water vapor performance. A scoring system is devised to quantitatively evaluate and rank the AR5 model performances, and this is applied to 30°N-30°S oceanic regions where the effects of diurnal variations are small. The organization of the paper is as follows: section 2 describes the AR4/AR5 models and their outputs used here; section 3 describes the A-Train datasets; section 4 compares model outputs, including differences between AR4 and AR5 model versions, and differences with the A-Train observations; and section 5 describes the model scoring system and gives performance results and model ranking based on this scoring system. We note that a number of other studies [*Li et al.*, 2005; *Su et al*., 2006; *Li et al*., 2007; *Waliser et al*., 2009; *Jiang et al.,* 2010; *Su et al.,* 2011] have used A-Train cloud and water vapor data to evaluate the performance of AR4 global circulation models.

2. AR4 and AR5 Climate models

We analyzed output from 12 AR4 and 19 AR5 models that, at the time of our analyses, had been submitted to the Program for Climate Model Diagnosis and Inter-comparison (PCMDI) Earth System Grid (ESG) [see http://pcmdi3.llnl.gov/esgcet/] Coupled Model

Intercomparison Project (CMIP). These models are listed in Table 1, along with horizontal resolutions and references. Fifteen AR5 models are coupled atmosphere-ocean models, while four (CCCMA am4, GFDL am3, GISS e2-h, and UKMO hadgem2-a) are atmosphere models. For comparisons and evaluations, we re-grid all model data to a standard grid of 144×91 (longitude×latitude) with 2.5° (longitude) × 2° (latitude) horizontal resolution and 40 pressure levels from the surface to 24 hPa, with intervals of 50 hPa in the middle troposphere and finer near the boundary layer and the tropopause. The model results then used for comparison with A-Train data are averages - using these gridded data - of multi-year monthly outputs from the "historical*" runs (or AMIP runs for some atmospheric GCMs) for AR5 models and the "20c3m" runs for AR4 models, which are defined as simulations of recent past climate [*Taylor et al.* 2011]. The multi-year model averages are 20-year (1980-2000) mean when accessing progressing from AR4 to AR5 (section 4); or 25-year (1980-2005) mean when comparing between AR5 and A-Train (section 5). The end year is due to the end of "historical" forcing in AR4 and AR5 runs.

The model output cloud parameters used in this study are *clivi, clwvi, cli,* and *clw* [See the PCMDI standard output document by K. Taylor, under "*Requested Variables*" at http://cmip-pcmdi.llnl.gov/cmip5/output_req.html?submenuheader=2#req_list]. The parameter *clivi* is the vertically-integrated ice water path (IWP), *clwvi* is the vertically-integrated cloud water path (CWP) that includes both IWP and liquid water path (LWP), *clw* is the cloud liquid water mass mixing ratio, and *cli* is the cloud ice mixing ratio. This naming convention sometimes causes confusion since LWP is obtained by subtracting *clivi* from *clwvi*, but LWC and IWC are obtained directly from *clw* and *cli*. However, *clwvi* output from the AR4 models BCCR bcm2 and CSIRO mk3, and from the AR5 models CSIRO mk3.6 and IPSL cm5a, are for liquid water only. The parameter *prw* is vertically-integrated water vapor (i.e., precipitable

water), and *hus* is specific humidity. Table 2 summarizes the model outputs and acronyms used here.

3. A-Train data

NASA's A-Train (Aqua, Aura, CloudSat and CALIPSO satellites) carries a suite of sensors that provide nearly-simultaneous and co-located measurements of multiple parameters that can be used for evaluating aspects of climate model performances. The measurements used in this study, summarized in Table 3 with their estimated uncertainties, are (a) water vapor (H_2O) profiles from the Atmospheric Infrared Sounder (AIRS) onboard Aqua launched in 2002, (b) water vapor paths (WVP) from the Advanced Microwave Scanning Radiometer for Earth-Observing-System (AMSR-E) on Aqua, (c) liquid/ice water paths (LWP/IWP) from the Moderate-resolution Imaging Spectroradiometer (MODIS) on Aqua, (d) upper tropospheric H_2O and ice water content (IWC) profiles from the Microwave Limb Sounder (MLS) on Aura launched in 2004, and (e) liquid water content (LWC) and IWC from CloudSat launched in 2006, and (f) IWC from CALIPSO also launched in 2006.

AIRS version 5, Level 3 H_2O product AIRX3STD is used (*Olsen et al.* 2007); it has spatial resolution of 50 km, but reported on 1°× 1° (longitude × latitude) grid. The useful altitude range is 1000 hPa to 300 hPa over ocean and 850 hPa to 300 hPa over land. The estimated uncertainty is 25% in the tropics, 30% at mid-latitudes, 50% at high latitudes and 30% globally averaged. The AIRS WVP over land is computed as the vertical integration of water vapor content from 850 hPa to 300 hPa and the AIRS WVP over ocean is the vertical integration from the 1000 hPa to 300 hPa.

AMSR-E Level 3 WVP data are used: the Version 5 ocean product (*JAEA*, 2005) downloaded from the Remote Sensing Systems website (http://www.remss.com) and reported on 0.25°×0.25° (longitude×latitude) grid. The AMSR-E WVP is expected to be slightly larger

over the land as AMSE-E measures the total water vapor content from the surface to the top of atmosphere. The AIRS science team has done a detailed comparison study of the WVP's between AMES-E and AIRS over the ocean, and found that the difference is no more than 5%.

MODIS daily IWP and LWP data are used: from the Collection 005 Level-3 MYD08-D3 product [*Hubanks et al.,* 2008] are generated by sub-sampling high resolution (1km), Level-2 swath product (MYD06). These data are binned at 1° × 1° (latitude × longitude) resolution. We note that the MODIS original IWP and LWP values are for cloudy scenes only, which were computed for each grid box as total retrieved IWP or LWP divided by number of successful cloud retrievals. For consistency with the gridded model data, we re-computed the MODIS original IWP and LWP to include both cloudy and clear sky scenes (by multiplying the original values by the cloud fractions for ice and liquid clouds, respectively). Thus the MODIS IWP and LWP used here are calculated as total retrieved IWP or LWP divided by number of both clear and successful cloud retrievals for each grid-box. The MODIS data uncertainties include the effects of baseline and particle size distribution (PSD) assumptions. In the absence of other information, we assume a factor of 2 as a realistic uncertainty estimate for MODIS IWP and LWP (*Steven Platnick, personal communications*), which is similar to the IWP and LWP uncertainties described below for MLS and CloudSat.

For MLS we use version 2.2 Level 2 [*Livesey et al.*, 2007] IWC and H_2O datasets, whose validations are described by *Read et al.* [2007] and *Wu et al.* [2008], respectively. These data have vertical resolution of ~3-4 km, and horizontal resolutions of ~7 km across-track and ~200-300 km along-track. The useful altitude ranges are from 215 hPa to 83 hPa for IWC, and pressure < 316 hPa for H_2O. The measurement uncertainties (including biases) for H_2O are 20% (215 hPa) to 10% (100 hPa) at tropics and mid-latitudes, and ~50% at high latitude (>60°N/S) (*Read et al.* 2007). For IWC, there is a factor of 2 uncertainty (*Wu et al.*, 2008),

which is mostly scaling uncertainty associated with the microphysics assumptions, e.g. Particle Size Distribution (PSD) assumption, in the MLS forward model used for retrievals. The MLS WVP is computed as the vertical integration of MLS H_2O from the 215 hPa to the top of atmosphere.

CloudSat IWP, LWP, IWC, and LWC data from the 2B-CWC-RO (version r04) dataset, whose retrievals are described by *Austin et al*. [2009], are used. These data have horizontal resolution of ~2.5 *km* along-track and ~1.4 *km* cross-track. The vertical resolution is ~480 *m*, oversampled to 240 *m*. One of the major uncertainties is that the retrieved IWC and LWC include some contributions from precipitating particles. Thus CloudSat IWC and LWC are likely overestimated. Profiles where precipitation was detected are removed by using the CloudSat 2C-PRECIP-COLUMN product (*Haynes et al.*, 2009), which flags precipitation (rain, snow, drizzle and graupel) for each IWC and LWC profile over the oceans. An average computed using no precipitation LWC or IWC profiles is called the *noPcp* value, while an average computed using all the IWC or LWC profiles is called the *Total* value. The *noPcp* values, as noted by *Eliasson et al.* (2011), inevitably have a low bias as all "floating" ice or liquid cloud particles associated with precipitation events are removed. Nevertheless, the range between *noPcp* and *Total* provides a reasonable estimate of the lower and upper uncertainty bounds on CloudSat IWC and LWC. Validation studies by *Heymsfield et al.* [2008], *Eriksson et al.* [2008], and *Wu et al.* [2009], indicate that the CloudSat retrieval error is likely within ~50%. Similar to the MLS IWC, the CloudSat IWC and LWC also have uncertainty due to the PSD assumption. CloudSat IWC/LWC estimated uncertainty is a factor of 2. Therefore, for the model comparisons, we use 0.5× the *noPcp* value as the low-end of the CloudSat uncertainty, and 2.0× the *Total* value as the high-end of the uncertainty. The cloud water content (CWC) is the sum of IWC and LWC.

CALIPSO IWC data from version 3.1 Level 2 L2-LIDAR-CPRO datasets are used. These data have horizontal resolution of 5 *km* along-track, ~1 *km* cross-track, and vertical resolution of 500 *m*. The uncertainty of CALIPSO IWC - including scaling error due to PSD - is assumed to be a factor of 2 (*Melody Avery, personal communications*).

All the A-Train datasets were put onto the same 144 (longtitude) × 91 (latitude) × 40 (pressure) grid as done for the model outputs. The A-Train multi-year means used in comparisons with the models, and for calculating the various model performance scores are averages of these gridded data over the following time periods: 5 years (August 2006 to July 2010) for CloudSat and CALIPSO; 8 years (October 2002 to September 2010) for AIRS and AMSR-E, 6 years (October 2002 to September 2008) for MODIS, and 7 years (September 2004 to August 2011) for MLS. Although the A-Train time periods do not overlap with those of the model outputs, no significant trends in clouds and water vapor are found in the model simulations and both the model and A-Train averages are thus expected to represent "recent past climate" for which our analyses are intended.

The A-Train satellites are sun-synchronous with equatorial crossings at ~1:30pm and ~1:30am, and this can cause sampling biases for parameters (e.g, IWC) that have diurnal variation. To reduce the effects of diurnal sampling bias, when quantitatively scoring the model performances we use A-Train and model data only from the tropics and subtropics (30°N to 30°S) and only over oceanic regions – where diurnal variations are much less than over land. To estimate the amount of residual diurnal bias between model and A-Train means, we took daily 3-hour IWC data from 3 models (NCAR, GFDL, and GEOS5, used by *Su et al*. 2011) and computed 30°N to 30°S oceanic means after interpolating to the MLS sampling times: differences with the model monthly means over tropical ocean were 1.5% for NCAR, 0.9% for GFDL, and 0.1% for GEOS5 (compared to up to 200% differences for non-oceanic

land regions). We thus estimate that diurnal variation introduces a bias of less than 2% in our 30°N to 30°S oceanic mean comparisons between model and observation, significantly smaller than the measurement uncertainties.

4. Comparisons of model outputs and A-Train observations

4.1 IWP, LWP, and WVP

Figure 1 shows the global, tropical (30°S-30°N), mid-latitude (30°N/S-60°N/S) and high-latitude (60°N/S-80°N/S) multi-year averages of IWP, LWP and WVP from AR4, AR5 and A-Train. As a major objective of this figure is to illustrate changes between the AR4 and AR5 outputs, we include only results from models for which both AR4 and AR5 outputs were available. Grey horizontal bands in the IWP and IWC panels show the global mean 'best estimate' range - the range between CloudSat *Total* and *noPcp* global means. The factor of 2 uncertainty limits for the global IWP and LWP best estimates are shown by dotted lines. Note that MODIS IWPs for all three zonal means, and the global mean, and are within the CloudSat grey band, supporting a 'best-estimate' interpretation for this band. However, MODIS measures IWP only in sunlight and its high-latitude mean does not include IWPs from the dryer polar winter. The MODIS global and mid-latitude mean LWPs are also within the grey band. The uncertainty limits of WVP global mean measurements, estimated as ±30% of the AIRS+MLS global mean WVP, are also shown by dotted lines. The AIRS+MLS WVPs are computed using the AIRS and MLS H_2O measurements both over land ($P \leq 850$ hPa) and over ocean. To facilitate the comparison between the models and AIRS+MLS, the model WVPs over land are also computed as the vertical integration of *hus* from 850 hPa to the top of atmosphere and the model WVPs over ocean are computed as the usual vertical integration from the surface to the top of atmosphere. The AMSR-E WVPs are the total water vapor content from the surface to the top of atmosphere, but over the ocean only.

4.1.1 IWP multi-year global and zonal means

The most notable change in AR4 to AR5 model outputs is the ~50% reduction of mid-latitude and high-latitude IWP from GISS e-h/e-r to e2-h/e2-r, seen in the top panel of Figure 1. This reduction results from two modifications in the GISS model ice cloud microphysics: (1) increasing the rate of conversion from cloud ice to snow; and (2) removing the influence of convectively-generated snow on the glaciations of lower super-cooled liquid cloud layers. These modifications take effect mostly over the mid and high latitudes. The tropical mean IWP in GISS e2-h/e2-r is increased by ~15% compared to e-h/e-r. Although still ~30% higher than the upper end of the A-Train best-estimate range, both GISS AR5 models produce IWP within the observational uncertainty, a significant improvement from the AR4 models.

Tropical IWP is notably increased from GFDL's AR4 cm2 to its AR5 cm3 model that addresses cloud-aerosol interaction and atmospheric chemistry issues that were not treated in cm2. Cloud particle concentrations in cm2 were specified as constants, whereas in cm3 they are related to droplet activation that depends on aerosol properties and vertical velocity [*Ming et al.*, 2006]. Also, interactive atmospheric chemistry is in cm3 instead of the specified chemical and aerosol concentrations in cm2. See *Donner et al.* [2011] for more information on the formulation of cm3 and the changes from cm2.

The AR5 models CCCMA canesm2, MIROC miroc5, and UKMO hadgem2 also show increases of global IWP from their AR4 counterparts. For the CCCMA model, its AR5 version differs substantially from the previous AR4 in its treatment of a number of physical processes. In particular, the CCCMA model now includes prognostic representations of stratiform clouds and aerosols and their direct and indirect effects on climate. In addition, treatments of radiative transfer, convection, and turbulent mixing were completely revised. However, exactly what caused the improvement between AR4 and AR5 is not known. The Japanese new AR5 model MIROC miroc5 employs an upgraded parameterization schemes. In

particular, treatment of clouds is substantially different from, and has larger degrees of the freedom than the previous AR4 model miroc3.2. For the UKMO hadgem2, a recent study by *Martin et al.* [2010] have shown significant improvements globally for the simulation of cloud amount and humidity compared to its predecessor hadgem1. This is particularly apparent in the tropics and results primarily from changes to the convection scheme. These changes include an "adaptive detrainment" parameterization [*Derbyshire et al*., 2011], exponential decay of convective cloud with a half-life of 2 hours, and removal of the depth criterion for shallow convection [*Gregory and Rowntree,* 1990].

Reductions of IWP in AR5 compared to AR4 are seen in BCCR noresm1, INM cm4 and NCAR cam5. The NCAR cam5 IWP output include floating snow ice [*Gettelman et al.* 2010a], but the model's mean IWPs are notably even smaller than the CloudSat *noPcp* values. IWPs from CNRM, CSIRO and IPSL models show little change between AR4 and AR5. The IPSL cm5a model is very similar of the previous IPSL cm4 model except for improvements of horizontal and vertical resolutions [*Dufresne et al.* 2011], whereas the changes made in the CNRM's and CSIRO's AR5 models have little effect on their IWPs.

Overall, of the 12 model pairs examined, 7 AR5 IWPs are within the CloudSat grey band, and 11 (all except INM cm4) are within the observational uncertainty limits. This is an improvement over AR4, where 6 models have IWPs within the grey band and 8 have IWPs within the uncertainty limits.

4.1.2 LWP multi-year global and zonal means

The middle panel of Figure 1 shows LWP, where increases from AR4 to AR5 model outputs are seen in BCCR noresm1, CCCMA canesm2, GISS e2-h and e2-r, INM cm4, and UKMO hadgem2. Reductions in LWPs from AR4 to AR5 are seen in CNRM cm5, CSIRO mk3.6, GFDL cm3, IPSL cm5a, NCAR cm5, and MIROC miroc5. Some of these changes in LWP are related to changes in cloud treatment in the model. For example, the CSIRO model

includes a simple treatment of sub-grid moisture variability, in which the width of sub-grid moisture distribution is parameterized via a prescribed critical relative humidity (RH_c) for onset of cloud formation [*Rotstayn*, 1997]. In mk3, RH_c was reduced between cloud base and top in convective columns when convection occurs. It has shown that such RH_c reduction leads to a large increase of LWP, whereas IWP is relatively insensitive to RH_c [*Rotstayn*, 1999]. This feature was removed in mk3.6, in which the RH_c is prescribed and no dependence on convection. This change explains the substantial decrease of LWP from mk3 to mk3.6, in conjunction with only a small change in IWP.

Global mean LWPs within the grey band are produced by 4 AR5 models: GDFL cm3, INM cm4, NCAR cam5, and UKMO hadgem2. Eleven AR5 models (all except BCCR noresm) have LWPs within the observational uncertainty. In contrast, only 2 AR4 models (GISS e-h and e-r) give global mean LWPs within the grey band, and 11 AR4 models (all except MIROC miroc3.2) have LWPs within the observational uncertainty.

4.1.3 WVP multi-year global and zonal means

The lower panel of Figure 1 shows WVP. Model differences are within ~10%, and changes from AR4 to AR5 are less than 5%. The differences between model and AIRS+MLS observation are less than ~15%, well within the 30% observational uncertainty. The difference between AIRS+MLS and AMSR-E are mainly due to the fact that AMSR-E WVPs do not include data over land, whereas the AIRS+MLS (and all models') WVPs are averaged using both data over oceans (pressure ≤ 1000 hPa) and data over lands (pressure ≤ 850 hPa).

4.1.4 IWP multi-year mean spatial distributions

Figure 2a shows the multi-year mean spatial distributions of IWP from the AR4 and AR5 models and from the A-Train. The significant changes in IWP from the AR4 to AR5 models in comparison with the observations, are:

- BCCR (bcm2 to noresm): Overall reduction in IWP results in low bias compared to the observations.
- CCCMA, cgcm3.1 to canesm2: Overall increase in IWP results in substantial improved agreement with observations.
- CNRM, cm3 to cm5: Very little change. Both models have morphology very similar to CloudSat observations. Their IWP values between the CloudSat *Total* and CloudSat *noPcp*.
- CSIRO, mk3 to CSIRO mk3.6: Slightly reduced IWP in the tropics results in a slight degradation in the agreement with observations.
- GFDL, cm2 to cm3: IWP increase in the tropics but decrease in the northern hemispheric storm tracks and southern mid and high latitudes, gives better agreement with observations in the tropics, but a low bias in the mid and high latitudes.
- GISS, e-r(h) to e2-r(h): Substantial reduction in mid and high latitude IWP, and increase in the tropics, result in better agreement with observations.
- INM, cm3 to cm4: IWP decrease in the equatorial eastern Pacific but increase over the mid-latitude storm tracks. The global mean is not significantly changed, but there is noticeable degradation in agreement with observations over the inter-tropical convergence zone (ITCZ).
- IPSL, cm4 to cm5a: Changes are very small, but IWP in the tropics is slightly reduced resulting in slightly improved agreement with observations there.
- MIROC, miroc3.2 to miroc5: IWP increased slightly over both the tropics and mid-latitudes, resulting slightly improved agreement with the observations.
- NCAR, ccsm3 to cam5: IWP reduced slightly over the oceans, but increased over the landmasses. There is no obvious improvement compared to observations.
- UKMO, hadgem1 to hadgem2-a: Slight increase in IWP in the tropics results in smaller low bias compared to observations; little change in the mid- and high latitudes.

Of the 12 AR5 models examined, comparisons with the observations indicate that 7 models show IWP improvements from AR4, 2 show little change, and 3 appear degraded.

4.1.5 LWP multi-year mean spatial distributions

Figure 2b shows the multi-year mean spatial distributions of LWP from the AR4 and AR5 models and from the A-Train. Comparisons with the CloudSat observations show that most

models have significant disagreement in the eastern Pacific subsidence region. Changes in LWP from AR4 to AR5 are:

- BCCR, bcm2 to noresm: Large increase in LWP leads to significant overestimate compared to observations, and worse performance than its older AR4 version.
- CCCMA, cgcm3.1 to canesm2: Large increase in LWP and the appearance of a "double ITCZ" in the equatorial Pacific result in poorer agreement with observations;
- CNRM, cm3 to cm5: No significant change, but except slightly reduced IWC results in slightly improved agreement with the observations.
- CSIRO, mk3 to mk3.6: Reduced LWP in mid-latitudes results in substantial improvement (in both amount and distribution) comparing to observations. Also notable is the improved simulation of clouds in the eastern Pacific subsidence region and the southern Indian Ocean west of Australia.
- GFDL, cm2 to cm3: Spatial patterns are similar, but magnitude of LWP is reduced, resulting in better agreement with observations. The morphology of LWP in the GFDL models is generally similar to the observations, but stratiform clouds near the coast of Peru are not captured well, especially in cm3.
- GISS, e-h(r) to e2-h(r): LWP increases, with more substantial increases in mid and high latitudes than in the tropics. Spatial distribution appears more zonal than the observations. It is not clear whether there is improvement compared to the observations.
- INM, cm3 to cm4: Slightly increased LWP results in better agreement with observations.
- MIROC, miroc3.2 to miroc5: Substantial reduction in LWP results in better agreement with observations.
- NCAR, ccsm3 to cam5: Substantial reduction in LWP results in better agreement with observations.
- UKMO, hadgem1 to hadgem2: Increased LWP results in better agreement with observations.

Of the 12 models examined, 8 show LWP improvements from AR4 to AR5, 2 show changes but no notable improvements, while 2 appear degraded, compared with observations.

4.1.6 WVP multi-year mean spatial distributions

Figure 2c shows the multi-year mean spatial distributions of WVP from the AR4 and AR5 models and from the A-Train. There is overall good agreement with the observations, and

model differences are small. Since the variability of WVP is dominated by lower-tropospheric water vapor, it is expected that the simulated lower tropospheric water vapor is similar among models, while large discrepancy may exist in the upper troposphere as we will discuss later.

4.2 Vertical Profiles of CWC, IWC and H_2O

Figure 3a shows the multi-year mean vertical profiles of CWC and IWC (upper panels) and H_2O (lower panels) from the 19 AR5 models and from the A-Train observations. The 'best estimated' CWC values from the CloudSat observations are indicated by the grey band between the CloudSat *noPcp* and *Total* values. Observational uncertainty limits are indicated by the dotted lines. There is a large spread among model CWC in all three latitude bands and globally. At 300 hPa, for example, the global mean CWC from GISS e2-r is more than 200× larger than from INM cm3. The differences between MLS and CALIPSO IWC in the upper troposphere ($P \leq 215$ hPa) are less than factor of 2, consistent with their estimated uncertainties. The modeled tropical CWCs range from ~3% to ~15× of the MLS observations in the upper troposphere. For mid-troposphere 700 hPa to 400 hPa, the modeled tropical CWCs are from ~30% to ~ 4× of the CloudSat *Total*. In lower troposphere, the modeled CWCs are ~ 40% to 2× of the CloudSat *Total*.

H_2O (lower panel of Figure 3a) differences among the models are within 20% in the mid- and lower troposphere, but more than 400% above ~200 hPa altitude. Model differences from the AIRS observations are small (< 10%) in the mid- and lower troposphere, but range from ~1% to ~200% of the MLS observations at 100 hPa.

Figure 3b shows the multi-year zonal means of CWC and H_2O as a function of latitude and height. Some major points to be noted from this figure are:

- The BCC csm1 and BCCR noresm model outputs are remarkably similar. This is not surprising since both models use similar cloud treatment. The cloud scheme in BCC csm1 is the same as that used in NCAR cam3 [*Boville et al.* 2006]. The atmospheric

component in the Norwegian Earth system model BCCR noresm is also modified from the NCAR model by updating the NCAR cam3's aerosol-cloud scheme [*Seland et al.* 2008; *Kirkevåg et al.* 2008; *Hoose et al.* 2009]. Both the BCC and BCCR models have large CWC amount in mid-layer clouds in the tropics, and low-level clouds at mid and high latitudes. However, their ice clouds have smaller IWC than CloudSat *Total*.

- The CCCMA models, am4 and canesm2, have nearly identical outputs with lower tropospheric CWC notably greater than the observations and have larger amounts of lower and mid-latitude CWC than the observations.
- The French CNRM cm5 and Australian CSIRO mk3.6 models have mid-latitude CWC similar to CCCMA models. However, the CNRM model has larger CWC than the CSIRO model in the tropics. The later has much smaller high-altitude tropical CWC compared to CloudSat *Total*, which may due to the exclusion of precipitation calculation of CWC in CSIRO mk3.6.
- The GFDL models, am3 and cm3, have the highest cloud top heights, which extend above 100 hPa, which is higher than the CloudSat observation. The two models also have substantial mid-tropospheric clouds (e.g. at ~600 hPa) near the tropics, which is also higher than the observation in terms of CWC amount.
- The GISS models, e2-h and e2-r, have morphology similar to CloudSat *Total*, but have larger amounts of IWC for the ice clouds.
- INM cm4 model has very small amounts of CWC relative to the observations.
- IPSL cm5a model produces large CWCs for mid-level and low level clouds at mid-latitudes. It simulates many clouds but very little high clouds in the tropics, compared with observations.
- NCAR cam5 has many low level clouds and very weak mid-level and high clouds, compared with observations.
- For the Japanese models, MIROC miroc4h and MRI cgcm3 have similar mid-latitude CWCs. They both have more CWC at mid-altitude and less CWC at high altitude, compared to CloudSat observations. MIROC miroc5 has smaller CWC at mid- and high altitudes compared to miroc4h and MRI cgcm3. The miroc5 model also appears to have substantial more CWCs at the northern subtropics than the southern subtropics.
- The three UKMO models, hadgem2-a, hadgem2-cc, and hadgem2-es, are very similar. Compared with observations, they produce too much mid-latitude CWC in the mid-

and lower troposphere, but too little CWC in the tropical upper tropospheric deep convective regions.

- All models produce similar zonal mean distributions of water vapor. The major differences with observations are in the upper troposphere where, for example, MIROC miroc5 produces less than 1/10 the amount of H_2O observed by MLS.

5. Quantitative evaluation of model performances

In this section we quantify the differences between model and A-Train multi-year means, and score the model performances compared to the observations. Only 30°S-30°N oceanic regions are considered in order to reduce diurnal sampling biases. It is in the tropics and subtropics that climate model performance is most critical for future climate prediction.

5.1 The scoring system

Model performance is evaluated with a system that scores how well each model multi-year mean reproduces the A-Train multi-year mean in terms of (1) spatial means, (2) spatial variances, and (3) spatial distributions. Our scoring system follows that of *Douglass et al.* [1999], *Waugh and Eyring* [2008], and *Gettelman et al.* [2010b], but with additional considerations of observational uncertainties.

We define the spatial mean scores G_m for IWC, LWC and H_2O as

$$G_m^{IWC,\ LWC} = \max\left[0, 1 - \frac{1}{n_g}\frac{\left|\ln(m_{mdl}^{IWC,\ LWC}) - \ln(m_{obs}^{IWC,\ LWC})\right|}{\ln \varepsilon_{m,obs}^{IWC,\ LWC}}\right], \quad (1)$$

$$G_m^{H_2O} = \max\left[0, 1 - \frac{1}{n_g}\frac{\left|m_{mdl}^{H_2O} - m_{obs}^{H_2O}\right|}{\varepsilon_{m,obs}^{H_2O}\, m_{obs}^{H_2O}}\right], \quad (2)$$

where m denotes the 30N-30S oceanic spatial mean, *mdl* denotes model value, *obs* denotes observational value, and $\varepsilon_{m,obs}$ is the fractional uncertainty of the observed spatial mean. The observed IWC and LWC spatial means have a factor of 2 uncertainty; hence $\varepsilon_{m,obs}^{IWC,LWC} = 2$. The H_2O observational uncertainties $\varepsilon_{m,obs}^{H_2O}$ are 0.1 at 100 hPa, 0.2 at 215 hPa, and 0.25 at 600 and

900 hPa. The scaling factor n_g is chosen to be 3, except for LWC at 900hPa where n_g = 4 is chosen to account for a greater uncertainty (50% of noPcp value) in LWC there. Due to the large range of values, the difference in logarithms is used for IWC and LWC. In this grading system, for example, a zero G_m score means: (1) for H_2O, the model-observation difference is greater than 3× the observational uncertainty, and (2) for IWC/LWC, the model value is either 8× greater (16× for 900hPa) or less than 1/8 (1/16 for 900hPa) the observational value.

Similarly, we define the spatial variance scores G_v as:

$$G_{\mathrm{v}}^{IWC,LWC} = \max\left[0, 1 - \frac{1}{n_g}\frac{\left|\ln\sigma_{mdl}^{IWC,LWC} - \ln\sigma_{obs}^{IWC,LWC}\right|}{\ln\varepsilon_{\mathrm{v},obs}^{IWC,LWC}}\right], \quad (3)$$

$$G_{\mathrm{v}}^{H_2O} = \max\left[0, 1 - \frac{1}{n_g}\frac{\left|\sigma_{mdl}^{H_2O} - \sigma_{obs}^{H_2O}\right|}{\varepsilon_{\mathrm{v},obs}^{H_2O}}\right], \quad (4)$$

where σ_{mdl} and σ_{obs} are the standard deviations from models and observations, respectively. The uncertainty of the observed spatial variance, $\varepsilon_{\mathrm{v},obs}$, is the same as for $\varepsilon_{m,obs}$ discussed above and the same n_g values are also used here.

For the spatial distribution performance, we simply use spatial correlations between model and observation as the scoring system:

$$G_c = \max\left[0, C_{mdl,obs}\right], \quad (5)$$

where $C_{mdl,obs}$ is the spatial correlation between the multi-year mean from a model and the multi-year mean from the A-Train.

5.2 Bi-variate metrics for H_2O and LWC/IWC

As water H_2O is strongly coupled with LWC/IWC, it is informative to simultaneously analyze the model performances for H_2O and LWC/IWC. This is particularly useful in the tropical tropopause layer (TTL) where the sum of IWC and H_2O is nearly constant [e.g. *Flury*

et al. 2011]. We thus use bi-variate metrics (BVC) in the following sections to simultaneously evaluate the model performances for H_2O and for LWC.

5.3 Model performances in regards to spatial means

Figure 4 shows scatter plots of H_2O versus IWC at 100 hPa and 215 hPa, and H_2O versus LWC at 600 and 900 hPa. Black dots, and horizontal and vertical lines, show the A-Train multi-year means; the grey area indicates the observational uncertainties. Colored dots/cycles are the multi-year means from the AR5 various models. Black open-cycles represent the "multi-model mean", which is a "virtual model" constructed by averaging all 19 models' outputs. Tables 4a and 4b give numerical values for the spatial means, and for the resulting performance scores discussed below.

100 hPa IWC and H_2O spatial mean performances: At 100 hPa, the GISS e2-r model receives the highest IWC score (G_m^{IWC} = 0.90), with BCCR noresm second (0.86) and GISS e2-h third (0.70). Next are MIROC miroc4h (0.64), CSIRO mk3.6 (0.45), IPSL cm5a (0.43), MRI cgcm3 (0.22), BCC csm1 (0.21) and UKMO hadgem2-a (0.05). The other 10 models received G_m^{IWC} scores of 0.0. For H_2O, the GFDL am3 model gives best performance at 100 hPa and receives the maximum possible score of $G_m^{H_2O}$ = 1.0. Next are BCCR noresm (0.97), CCCMA canesm2 (0.92), UKMO hadgem2-cc (0.91), GFDL cm3 (0.84), CCCMA am4 (0.78), and NCAR cam5 (0.65). Models receiving $G_m^{H_2O}$ scores in the range 0.31-0.57 are UKMO hadgem2-es (0.57), BCC csm1 (0.47), UKMO hadgem2-a (0.42) and MIRCO miroc4h (0.31). The remaining seven models receive $G_m^{H_2O}$ = 0.0.

215 hPa IWC and H_2O spatial mean performances: At 215 hPa, the best overall performance is by the three UKMO models as well as CNRM cm5, which give both IWC and H_2O approximately within the measurement uncertainties, earning them G_m^{iwc} scores in the

0.62-0.77 range for 215 hPa IWC and G_m^{h2o} scores in the 0.79-0.92 range for 215 hPa H_2O. BCC and INM models produce accurate 215 hPa H_2O, with $G_m^{H_2O} = 0.99$ and 1.0, respectively, but produce much too little 215 hPa IWC (far outside the observational uncertainty) giving them low scores of $G_m^{IWC} = 0.21$ and 0.0, respectively. The two CCCMA models and the IPSL model give the best agreement with observed 215 hPa IWC ($G_m^{IWC} \geq 0.98$) but produce far too much 215 hPa H_2O ($G_m^{H_2O} = 0.0$ for CCCMA models and 0.33 for IPSL cm5a). MIROC miroc4h and MRI cgcm3 also receive good scores (G_m^{IWC}=0.88) for 215 hPa IWC, but produce too much 215 H_2O ($G_m^{H_2O} < 0.14$). MIROC miroc5 has relatively good performance (G_m^{IWC}=0.67 and $G_m^{H_2O}$=0.66) compared to MIROC miroc4h. Both GFDL and both GISS models, as well as CSIRO mk3.6 also give far too much 215 hPa H_2O ($G_m^{H_2O}$ in the 0.0-0.16 range), and too much 215 hPa IWC ($G_m^{IWC} \leq 0.5$). The NCAR model produces 215 hPa IWC within the observational uncertainty ($G_m^{IWC} = 0.73$), but gives 215 hPa H_2O slightly outside the uncertainty ($G_m^{H_2O} = 0.55$). The BCCR noresm gives 215 hPa IWC slightly outside the uncertainty ($G_m^{IWC} = 57$), but its 215 hPa H_2O is further outside the uncertainty ($G_m^{H_2O} = 0.44$).

600 hPa LWC and H_2O spatial mean performances: At 600 hPa, model LWC ranges from 0.9 mg/m^3 (NCAR cam5) to 10.9 mg/m^3 (MRI cgcm3), compared to the Cloudsat value of 2.8 mg/m^3 (with uncertainty range from 1.3 to 5.6 mg/m^3). The three UKMO models and the Australian model CSIRO mk3.6 all receive excellent G_m^{LWC} scores (0.97 for hadgem2-a and hadgem2-cc, 0.99 for hadgem2-es, and 1.0 for CSIRO mk3.6). GISS and INM models receive good G_m^{IWC} scores (0.70s). The two CCCMA models, GFDL models, and IPSL cm5a have G_m^{IWC} scores in the 0.60s. Other models have G_m^{LWC} in the 0.40s or below. All models

perform well for 600 hPa H_2O, with differences from observations less than 20% - within the 25% observational uncertainty. All models receive $G_m^{H_2O}$ scores higher than 0.6 at 600 hPa.

900 hPa LWC and H_2O spatial mean performances: At 900 hPa, model LWCs range from 4.53 mg/m^3 (INM cm4) to 48.2 mg/m^3 (MIROC miroc4h) and are all within the CloudSat observational uncertainty. Scores are $G_m^{LWC} > 0.7$ or better, except for INM cm4 with $G_m^{LWC} = 0.39$, due to its LWC being smaller than the CloudSat *noPcp* value. All models perform well for 900 hPa H_2O, with scores $G_m^{H_2O} > 0.7$ or better.

In Figure 4, the "multi-model mean" at 900 hPa falls within the grey area, and it slightly over produces IWC and thus locates just outside the edge of the grey area at 600 hPa. At upper troposphere, the "multi-model mean" has excessive H_2O at 215 hPa and too large IWC at 100 hPa. As most models have high bias of 215 hPa H_2O, the "multi-model mean" is also biased high. At 100 hPa, although most models have low bias of IWC, the extremely large values of IWC from the two GDFL models make the "multi-model mean" IWC biased higher than the observation.

It is worth noting that there is no apparent correlation between model scores for IWC/LWC and for H_2O. Some models (e.g. CCCMA am4/canesm2 at 215hPa) simulate IWC/LWC extremely well compared to the observations, but give H_2O that is substantially different from observations. Other models (e.g. GFDL am3/cm3 at 100 hPa) simulate H_2O extremely well, but give IWC/LWC that is significantly different from the observations. This points to the need for developing more accurate and consistent model representations for physical processes jointly affecting clouds and water vapor.

5.4 Model performances in regards to spatial variations

We now examine the degree to which the spatial variations in the multi-year means from the AR5 models reproduce the spatial variations in the multi-year means from the A-Train

observations over 30ºS-30ºN oceanic regions. Tables 5a and 5b give numerical values for the spatial variance (standard deviation) and the resulting spatial variance scores. Tables 6a and 6b give numerical values for the spatial correlation and the resulting spatial correlation scores. Subsections below discuss the model performances at each of the 4 vertical levels. To succinctly illustrate the performances we use Taylor diagrams [*Taylor* 2001] in which spatial correlation, centered root-mean-square-differences, and amplitudes of spatial variations (represented by their standard deviations) are displayed simultaneously in a compact format. It should be noted that the means of the fields are removed before computing the statistics for the Taylor diagrams - so these diagrams do not provide information about the mean differences, but the mean differences are quantified in section 5.3 above.

Figure 5 gives Taylor diagrams for H_2O at 100, 215, 600 and 900 hPa, for IWC at 100 and 215 hPa, and for LWC at 600 and 900 hPa. Results are shown for all the 19 AR5 models that produce vertical profiles of H_2O, IWC and LWC. The Taylor diagram position of the symbol for each model quantifies how closely that model simulates the variation in the observed field. The centered root-mean-square difference between the modeled and observed field is proportional to the distance (green contours) between the point for that model and the point at unity value on the horizontal axis (black dot representing the normalized standard deviation of the observed field). The standard deviation of the model field itself is proportional to the radial distance between the point for that model and the origin. The coefficient of correlation between the modeled and observed fields is non-linearly related to the clockwise angle from the vertical axis, with correlation coefficient values indicated along the outer arc. It should be noted that the reason this 2-dimensional figure can represent these three different statistics simultaneously is that the three statistics are not independent of each other.

5.4.1 Model spatial variation performances at 100 hPa

Of the four vertical levels examined, differences among models - and differences between models and observation - have the largest spread at 100 hPa.

100 hPa IWC spatial variation performance: Best model performance for 100 hPa IWC spatial variance is by BCCR noresm (0.86 score, 0.74 standard deviation), GISS e2-h (0.72 score, 1.77 standard deviation), and MIROC miroc4h (0.71 score, 1.83 standard deviation). Best performance for spatial correlation is by MIROC microc4h (0.85 score), NCAR cam5 (0.84 score), CCCMA am4 (0.83 score), and GFDL am3 (0.82 score). The two GISS models produce spatial correlations with observations of ~0.25 and receive G_c^{IWC} = scores of ~0.25. BCCR noresm and IPSL cm5a give similar spatial correlations (~0.6) and receive ~0.6 G_c^{IWC} scores. IPSL cm5a, however, produces standard deviation of only 0.33 and receives 0.47 G_v^{IWC} score. Most other models also give correlations of ~0.6-0.7, but produce even smaller standard deviations (< 0.2) and receive G_v^{IWC} scores in the 0.0-0.28 range. GFDL am3/cm3 have the largest RMS differences (27×/16×) with observations, and the most discrepant standard deviations (28×/17×) and receive 0.0 for G_v^{IWC} score; their spatial correlation performance, however, is good (G_c^{IWC} scores 0.82/0.75). With 15 of the 18 models scoring poorly for spatial variance ($G_v^{IWC} \leq 0.5$), it is clear that simulation of tropical tropopause layer cloud is a challenging area.

100 hPa H_2O spatial variation performance: BCCR noresm, as it did for 100 hPa IWC, produces spatial variance (1.06 standard deviation) in closest agreement with the MLS observation and receives score $G_v^{H_2O}$ = 0.81. The second and third closest agreements with MLS observation are achieved by UKNO hadgem2-cc and CNRM cm5, receiving scores $G_v^{H_2O}$ = 0.73 and 0.62, respectively. All other models receive $G_v^{H_2O}$ scores of 0.1 or less, with

IPSL cm5a producing too little spatial variation and all other models producing too much. It should be noted that, because the MLS uncertainty here is only 10%, any model producing 100 hPa H_2O spatial standard deviation that differs from the MLS value by ≥ 30% receives $G_{\mathrm{v}}^{H_2O}$ = 0.0. For spatial correlation, BCCR noresm produces the smallest positive correlation (0.038) and receives $G_c^{H_2O}$ = 0.04. CNRM cm5 and UKMO hadgem2-cc receive good $G_c^{H_2O}$ scores of 0.81 and 0.87, respectively. BCC csm1, GFDL am3/cm3, MIROC miroc5, NCAR cam5, and UKMO hadgem2-a/hadgem2-es also have good spatial correlation with MLS 100 hPa H_2O ($G_c^{H_2O}$ scores of ~0.70s and 0.80s) - and also have small RMS differences - but, as noted above, they all receive very low spatial variance scores. CCCMA am4/canesm2 and GISS e2-h/e2-r give negative spatial correlations with the observation and therefore receive $G_c^{H_2O}$ = 0.0 score. All other models receive correlation scores in the 0.40s-0.50s range.

5.4.2 Model spatial variation performances at 215 hPa

<u>215 hPa IWC spatial variation performances:</u> At 215 hPa, CCCMA am4/canesm2, IPSL cm5a, and MIROC microc4h models produce IWC spatial variances near that observed by MLS, and receive G_{v}^{IWC} scores of 0.93/0.96/0.97, 0.90 and 0.95, respectively. Most other models produce less spatial variance than observed, and most of these receive G_{v}^{IWC} scores in the ~0.5-0.8 range - but BCC csm1, BCCR noresm, and INM cm4 produce spatial variance that is only 0.09, 0.24, and 0.02, respectively, of that observed and receive respective scores of only 0.0, 0.32, and 0.0. Four models produce spatial variance substantially larger than observed: GFDL am3/cm3 produce 2.9/2.6× more (and receive scores 0.49/0.55), GISS e2-h/e2-r produce 10.1/10.7× more (and receive 0.0 scores) and have the largest RMS differences (9.5/10) with the observations. In regards to spatial correlation, 18 of the 19 models receive G_c^{IWC} scores in the 0.60s-0.90s range; UKMO hadgem2-cc/hadgem2-es and GFDL am3 have

the best (and almost equal) performance in this category. INM cm4 produces relatively low correlation of 0.49 and receives $G_c^{IWC} = 0.49$ score. The high spatial correlation scores of most models for 215 hPa IWC indicate that most produce reasonably-accurate locations for deep convection. But the diversity of spatial variance scores indicates discrepancies among models (and with observations) on the intensity of this convection. Note that the 215 hPa IWC spatial variances from each model are closely related to the corresponding means (Figure 4a and Table 4a), and that both are indicative of convection intensity.

215 hPa H_2O spatial variation performances: GISS e2-r produces 215 hPa H_2O spatial variance closest to that observed by MLS and receives $G_v^{H_2O} = 0.99$ score. BCC csm1 produces too little variance (0.48 standard deviation; 0.13 score), the two CCCMA models am4/canesm2 produce too much variance (1.36/1.46 standard deviation; 0.40/0.24 scores), and CSIRO mk3.6 and GFDL am3 also produce too much variance (1.43 standard deviation, 0.29 score; 1.54 standard deviation, 0.10 score; respectively). UKMO hadgem2-cc has too little spatial variance (0.64 standard deviation, 0.41 score). Most other models receive $G_v^{H_2O}$ scores spread in the 0.5-0.9 range. All models produce 215 hPa H_2O good spatial correlations in the 0.74-0.94 range, another indication that most models produce reasonably-accurate locations for deep convection.

5.4.3 Model spatial variation performances at 600 hPa

600 hPa LWC spatial variation performances. At 600 hPa, the best performance for LWC spatial variance is by CSIRO mk3.6 and the three UKMO models with standard deviations (~1) very close to those observed and small (~0.8) RMS differences; they receive spatial variance scores of $G_v^{LWC} = 0.96$ for CSIRO mk3.6, 0.97 for UKMO hadgem2-a, 1.0 for UKMO hadgem2-cc, and 0.99 for UKMO hadgem2-es. The spatial correlations for these models are in the 0.60s-0.70s range. Good spatial correlation performance is also had by BCC

csm1, BCCR noresm, CNRM cm5, GFDL am3/cm3, IPSL cm5a, MRI cgcm3, and NIES miroc4h - with G_c^{LWC} scores above 0.6. GFDL am3 has the highest spatial correlation (0.81). The GFDL am3/cm3 standard deviations, however, are about twice that observed, and they receive G_v^{LWC} scores of 064/0.68. BCC csm1, BCCR noresm, CCCMA am4/canesm2, CNRM cm5, GISS e2-h/e2-r, IPSL cm5a, MIROC microc4h/miroc5, and MRI cgcm3 have LWC standard deviations about 3× to 4× than observed, giving them standard deviation scores in the 0.30s-0.40s range. NCAR cam5 and INM cm4 have moderate correlation (with G_c^{LWC} scores in the 0.50s), relatively small RMS differences (~0.8), with spatial variance about half that observed (scores G_v^{LWC} of 0.62 and 0.74, respectively). GISS e2-h and e2-r produce weakly negative correlations (−0.029 and −0.036 respectively), and receive G_c^{LWC} = 0.0. The two CCCMA models am4/canems2 produce relatively low correlations, and receive 0.37 G_c^{LWC} score. Overall, 11 of 19 the models receive low (< 0.5) LWC spatial variance scores, and 6 models receive low (< 0.6) LWC spatial correlation scores. Improvements in the representation of mid-level clouds are certainly needed in many models.

600 hPa H_20 spatial variation performances: At 600 hPa, all models perform reasonably well for H_2O, with spatial correlations above 0.8 and standard deviations close to that observed. All models receive $G_c^{H_2O}$ correlation scores in the 0.80s or 0.90s, with highest received by NCAR cam5 (0.97) and GFDL am3 (0.98). Highest $G_v^{H_2O}$ variance scores are received by UKMO hadgem2-a (0.99), MRI cgcm3 (0.98) and GFDL cm3 (0.98). Most models receive $G_v^{H_2O}$ scores of 0.7-0.9, with slightly lower received by BCC csm1 (0.56), GFDL am3 (0.65), and GISS e2-h (0.67).

5.4.4 Model spatial variation performances at 900 hPa

900 hPa LWC spatial variation performances: At 900 hPa, as seen from the bottom left panel of Figure 5a, all models produce less LWC spatial variation than observed by CloudSat. CSIRO mk3.6, CCCMA canesm and MIROC miroc4h are closest to the observed amount of spatial variance and receive scores G_{v}^{LWC} in the 0.90s. The spatial patterns for the CCCMA and MIROC models, however, have relatively low correlation observed and receive spatial correlation scores of $G_c^{LWC} < 0.5$, while the correlation score for CSIRO mk3.6 is the highest at 0.75. The INM model produces least spatial variance and receives the low score $G_{\mathrm{v}}^{LWC} =$ 0.07; it and there other models - BCC csm1, CNRM cm5, and MRI cgcm3 - give very low spatial correlation with observation, and receive scores in 0.10s-0.20s. The GFDL am3 model has the second largest spatial correlation and receives $G_c^{LWC} = 0.73$. All models have RMS differences with observations (normalized to the standard deviation in the observed field) between ~0.75 and ~1.0.

900 hPa H_2O spatial variation performances: For 900 hPa H_2O, all models have RMS differences with observations of 0.2-0.4, correlations of 0.89-0.96, and variances that are – at most - only slightly different. Their scores for standard deviation and correlation are at $G_{\mathrm{v}}^{H_2O} \geq$ 0.74 and $G_c^{H_2O} \geq 0.89$. That the models reproduce boundary layer H_2O distribution well is not surprising, and reflects the tight constraint on boundary layer H_2O by sea surface temperature (SST), which is specified or closely matched to the observed SST.

As the "multi-model mean" inherently smooths out individual models' spatial variations, it is not surprising the spatial variances of the "multi-model mean" are generally closer to the observations than individual models. The spatial correlations of "multi-model mean" are also the highest among all models.

5.5 Overall summary of model performance scores

Figure 6 gives an overall summary of all 19 models' performances in a color-coded display of each model's spatial mean, spatial variance and spatial correlation scores for all three parameters (H_2O, IWC and LWC) and all four pressure levels examined here. Although the score values are not directly comparable between IWC/LWC (clouds) and H_2O (water vapor), we find that – at all 4 pressure levels - most models simulate water vapor better than clouds. All models receive poorer scores in the tropical tropopause layer and upper troposphere than they do in the middle troposphere and boundary layer. The model H_2O and IWC at 100 and 215 hPa vary greatly from model to model, indicating the large differences (and, thus, overall uncertainty) in the various parameterizations and microphysics for processes affecting high-altitude clouds.

For spatial means, most models have better scores in both LWC and H_2O at 900 hPa (boundary layer) and 600 hPa (middle troposphere) than at 215 (upper troposphere) and 100 hPa (tropical tropopause layer). The scores for LWC at 215 and 100 hPa are also generally better than those for IWC at 215 and 100 hPa. Besides the "multi-model mean", the three UKMO models appear best when considering spatial mean performance over all vertical levels, while the BCCR model appears best in the upper troposphere.

For spatial variability, it is clear that models generally simulate 600 and 900 hPa H_2O (water vapor) better than LWC (clouds). Most models do not well simulate the observed variability of IWC (clouds) at 215 and 100 hPa. An interesting result is the better scores for correlation than for variance at 215 and 100 hPa, indicating that models generally simulate upper tropospheric cloud and water vapor spatial patterns (which are connected to regions of deep convection) better than they simulate the amount of spatial variation. Spatial patterns of low and mid clouds are not universally well simulated.

The "multi-model mean" exhibits relatively superior performance in all aspects of metrics in Figure 6, except its score for the 215 hPa mean H_2O is below 0.5. The low score for 215 hPa H_2O reflects the fact that most models have high bias of 215 hPa spatial mean H_2O compared to the observation. On the other hand, both high and low biases exist for other quantities in the models, thus the "multi-model mean" effectively averages out the biases and achieve a better performance than many individual models.

5.6 Overall scores and rankings

To obtain an overall performance score for each model at each pressure level, we simply average its scores for all three variables (H_2O, IWC, LWC), and all three categories (spatial mean, spatial variance, spatial correlation) at each pressure level. Table 7 gives these overall 'pressure-level' scores, and performance rankings in terms of this score, for each model. To obtain a single overall performance score for each model, we simply average the overall scores for the four pressure levels. This resulting overall score for each model is given in Table 8, along with each model's rank in terms of this score.

UKMO hadgem2-a/hadgem2-cc models tie for the highest overall score (0.73) and UKMO hadgem2-es has the second-highest score (0.71). Two of the UKMO models hadgem2-a/hadgen2-es also have the highest 600 hPa score (0.91), and another UKMO model hadgem2-cc ranked second best at both 100 hPa and 600 hPa. BCCR noresm has the third-highest overall score (0.70) and the highest 100 hPa score (0.69). In fourth place and fifth place are MIROC mioc4h (0.69) and IPSL cm5a (0.66), of which, IPSL cm5a also has the highest 215 hPa score (0.79). Tied at sixth highest overall score are CSIRO mk3.6 (0.65) and NCAR cam5 (0.65), with CSIRO having the highest 900hPa score (0.92). Tied at seventh are GFDL am3 and cm3 (0.64), with GFDL am3 also having the second highest 900 hPa score (0.86). Also tied at eighth place in overall score are CCCMA am4 and MIROC miroc5 (0.62), and both CCCMA canesm2 and CNRM cm5 follow closely at ninth (0.61). BCC csm1, GISS

e2-h/e2-r, INM cm4 and MRI cgcm3 received relatively lower overall scores although some aspects of their performances are quite good; for example, BCC csm1 performs relatively good at 100 hPa comparing to most other models, GISS e2-h/e2-r receive good 900 hPa score, INM cm4 performs well at 600 hPa, and MRI receives relatively good 215 hPa score.

Interestingly, the overall score for the "multi-model mean" turns out to be the best (0.78) among all models. This may be coincidental, but it is comforting as the use of multi-model ensembles in climate projections is a common practice and the "multi-model mean" is generally perceived as closer to the "truth" than any single model alone, as found in previous model evaluation studies [e.g. *Gleckler et al.* 2008].

6. Conclusions

Using A-Train observations, we have assessed the multi-year mean simulations of cloud and water vapor by IPCC AR4 and AR5 models. For clouds, apparent improvements in model simulations of IWP are identified in 7 models (CCCMA am4 and canesm2, GFDL cm3, GISS e2-h and e2-r, IPSL cm5a and MIROC miroc5). For LWP, improvements are found in 8 AR5 models (CNRM cm5, CSIRO mk3.6, GFDL cm3, INM cm4, NCAR cam5, MIROC miroc5, UKMO hadgem1 and hadgem2), comparing to their previous AR4 versions. For water vapor, changes in WVP from AR4 to AR5 are relatively insignificant.

We also examined vertical structure of CWC and H_2O produced by 19 AR5 models. The largest spread among models and their differences from A-train observations are at upper troposphere level.

We develop a grading scheme to quantitatively evaluate model performance in simulating clouds and water vapors at different vertical levels (from boundary layer to tropopause) in terms of spatial mean, correlation and standard deviation. Overall, we find water vapor is generally better simulated than clouds. Boundary layer water vapor is the best simulated, because of the strong constraint on boundary layer water vapor by SST. Tropopause layer

water vapor is very poorly represented. For spatial mean, upper troposphere ice cloud is worse simulated than lower and middle troposphere liquid cloud. For spatial correlation, model simulated clouds and water vapor at 215 hPa are better represented than boundary layer clouds. Spatial variances of clouds at all levels are poorly simulated, compared to A-Train observations.

Considering all grades equally weighted, the overall average grades of all models show that the UKMO hadgem2-a and hadgem2-cc perform the best, followed by UKMO hadgem2-es, BCCR noresm, MIROC miroc4h, IPSL cm5a, tied CSIRO mk3.6 and NCAR cam5, tied GFDL am4/cm4, tied CCCMA am4 and MIROC miroc5, and tied CCCMA canesm2 and CNRM cm5, etc. The "multi-model mean" also has the best overall performance.

We recognize that our scoring scheme is simplified; it nevertheless provides a quantitative measure of relative skills of current models in simulating clouds and water vapor.

Acknowledgments. The NASA ROSES10 AST and COUND programs fund this project. The authors acknowledge the support by the Climate Science Center at the Jet Propulsion Laboratory, California Institute of Technology, sponsored by NASA. We thank helpful discussion and comments from Stephen Platnick of the MODIS team, Melody Avery of the CALIPSO team, and the two CCCMA internal reviewers who provided constructive comments. We are very thankful to our colleagues from climate modeling centers across the globe, including BCC, BCCR, NCC, CCCMA, CNRM, QCCCE, CSIRO, GFDL, GISS, INM, IPSL, MIROC, MRI, NCAR, the University of Tokyo, and UKMO.

References

Arora, V. K., J. F. Scinocca, G. J. Boer, J. R. Christian, K. L. Denman, G. M. Flato, V. V. Kharin, W. G. Lee, and W. J. Merryfield (2011), Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases, *Geophys. Res. Lett.*, 38, L05805, doi:10.1029/2010GL046270.

Arakawa, A., and W. H. Schubert (1974), Interaction of a cumulus cloud ensemble with large-scale environment, *J. Atmos. Sci.*, 31 (3), 674-701.

Bodas‐Salcedo, A., M. J. Webb, M. E. Brooks, M. A. Ringer, K. D. Williams, S. F. Milton, and D. R. Wilson (2008), Evaluating cloud systems in the Met Office global forecast model using simulated CloudSat radar reflectivities, *J. Geophys. Res.*, 113, D00A13, doi:10.1029/2007JD009620.

Bony, S., et al. (2006), How well do we understand and evaluate climate change feedback processes?, *J. Clim.*, 19, 3445-3482.

Boville, B.A., P.J. Rasch, J.J. Hack, and J.R. McCaa (2006), Representation of clouds and precipitation processes in the Community Atmospheric Model version 3 (CAM3), *J. Climate*, 19, 2184-2198.

Cess, R. D., et al, (1990): Intercomparison and interpretation of climate feedback processes in 19 atmospheric general circulation models. *J. Geophys. Res*., 95, 16,601-16,615.

Cess, R. D., et al., (1996): Cloud feedback in atmospheric general circulation models: An update. *J. Geophys. Res*., 101, 12,791-12,794.

Collins, W. J. et al. (2011): Development and evaluation of an Earth-system model – HadGEM2, *Geosci. Model Dev. Discuss*., 4, 997-1062, doi:10.5194/gmdd-4-997-2011.

Derbyshire, S. H., Maidens, A. V., Milton, S. F., Stratton, R. A. and Willett, M. R. (2011): Adaptive detrainment in a convective parametrization. *Q. J. R. Met. Soc*., 137, 1856–1871, doi:10.1002/qj.875.

Diansky N.A., Bagno A.V., Zalesny V.B. (2002), Sigma model of global ocean circulation and its sensitivity to variations in wind stress. Izv. *Atmos. Ocean. Phys.* (Engl. Transl.), V.38, No. 5, pp. 477-494.

Diansky N.A., Volodin E.M, (2002), Simulation of present-day climate with a coupled Atmosphere-ocean general circulation model. Izv. *Atmos. Ocean. Phys.* (Engl. Transl.), V.38, No. 6, pp. 732-747.

Donner, L.J., and Co-Authors, (2011): The dynamical core, physical parameterizations, and basic simulation characteristics of the atmospheric component of the GFDL coupled model CM3, *J. Climate*, 24, 3484-3519, doi: 10.1175/2011JCLI3955.1.

Dufresne, J-L, and Co-Authors (2011), Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5. Submitted to *Clim. Dynamics*.

Eaton, B. (2011), User's Guide to the Community Atmosphere Model CAM-5.1, available on line at: http://www.cesm.ucar.edu/models/cesm1.0/cam/docs/ug5_1/book1.html.

Eriksson, P., et al. (2008), Comparison between early Odin-SMR, Aura MLS and CloudSat retrievals of cloud ice mass in the upper tropical troposphere, *Atmos. Chem. Phys.*, 8(7), 1937-1948.

Flury, T., D.L. Wu, and W.G. Read (2011), Correlation among cirrus ice content, water vapor and temperature in the TTL as observed by CALIPSO and Aura/MLS, *Atmos. Chem. Phys.*, in press.

Gettelman, A., X. Liu, S. J. Ghan, H. Morrison, S. Park, A. J. Conley, S. A. Klein, J. Boyle, D. L. Mitchell, and J.-L. F. Li (2010a), Global simulations of ice nucleation and ice supersaturation with an improved cloud scheme in the Community Atmosphere Model, *J. Geophys. Res.*, 115, D18216, doi:10.1029/2009JD013797.

Gettelman, A. et al. (2010b), Multimodel assessment of the upper troposphere and lower stratosphere: Tropics and global trends, *J. Geophys. Res.*, 115, D00M08, doi:10.1029/2009JD013638.

GFDL-AMDT (Geophysical Fluid Dynamics Laboratory Global Atmosphere Model Development Team), 2004: The new GFDL global atmosphere and land model AM2-LM2: Evaluation with prescribed SST simulations. *J. Climate*, 17, 4641-4673.

Gleckler, P.J., K.E. Taylor, and C. Doutriaux (2008), Performance Metrics for Climate Models, *J. Geophys. Res*., 113, D06104, doi:10.1029/2007JD008972.

Gregory, D. and Rowntree, P. R. (1990): A mass flux convection scheme with representation of cloud ensemble characteristics and stability dependent closure, *Mon. Wea. Rev.,* 118, 1483-1506.

HDL (The HadGEM2 Development Team (2011): The HadGEM2 family of Met Office Unified Model climate configurations, *Geosci. Model Dev*., 4, 723-757, doi:10.5194/gmd-4-723-2011

Hasumi, H. and S. Emori, eds., (2004), K-1 model developers: K-1 coupled model (MIROC) description, K-1 technical report, 1, H. Hasumi and S. Emori (eds.), Center for Climate System Research, University of Tokyo, 34pp.

Haynes, J. M., R. T. Marchand, Z. Luo, A. Bodas-Sakedo, and G. Stephens (2007), A multi-purpose radar simulation package: QuickBeam, *Bull. Am. Meteorol. Soc.*, 88(11), 1723-1727, doi: 10.1175/BAMS-88-11-1723.

Haynes, J. M., T.S. L'Ecuyer, G.L. Stephens, S.D. Miller, C. Mitrescu, N.B. Wood, and S. Tanelli: Rainfall retrieval over the ocean with spaceborne W-band radar, *J. Geophys. Res*., 114, D00A22, doi:10.1029/2008JD009973, 2009.

Heymsfield, A. J., et al. (2008), testing IWC retrieval methods using radar and ancillary measurements with in situ data, J. Appl. Meteorol. Climatol., 47(1), 135-163.

Hoose, C., J.E. Kristjansson, T. Iversen, A. Kirkevåg, Ø. Seland, and A. Gettelman (2009), Constraining cloud droplet number concentration in GCMs suppresses the aerosol indirect effect, *Geophys. Res. Lett*., 36, L12807, doi:10.1029/2009GL038568.

Hourdin F., M-A Foujols, F. Codron, V. Guemas, J-L Dufresne, S. Bony, S. Denvil, L.Guez, F. Lott, J. Ghattas, P. Braconnot, O. Marti, Y. Meurdesoif, L. Bopp, (2011), Climate and

sensitivity of the IPSL-CM5A coupled model: impact of the LMDZ atmospheric grid configuration. Submitted to *Clim. Dynamics*.

Hubanks, P. A., M. D. King, S. A. Platnick, and R. A. Pincus (2008), MODIS Atmospheric L3 Gridded Product Algorithm Theoretical Basis Document, in *MODIS Algorithm Theoretical Basis Document No. ATBD-MOD-30 for Level-3 Global Gridded Atmosphere Products (08_D3, 08_E3, 08_M3)*, Goddard Space Flight Center, Greenbelt, MD.

Jiang, J. H., and C. Zhai (2011), Comments to "Comparisons of satellite liquid water estimates with ECMWF and GMAO analyses, 20th century IPCC AR4 climate simulations, and GCM simulations" by Li et al., *Geophys. Res. Lett.*, to be submitted.

Jiang, J.H., H. Su, S. Pawson, H.C. Liu, W. Read, J.W. Waters, M. Santee, D.L. Wu, M. Schwartz, N. Livesey, A. Lambert, R. Fuller, and J.N. Lee (2010), Five-year (2004-2009) Observations of Upper Tropospheric Water Vapor and Cloud Ice from MLS and Comparisons with GEOS-5 analyses, *J. Geophys. Res.* 115, D15103, doi:10.1029 /2009JD013256

Johns, T.C. et al. (2006): The new Hadley Centre climate model HadGEM1: Evaluation of coupled simulations. *J. Climate*, 19, 1327-1353.

Jones, C. D. et al (2011): The HadGEM2-ES implementation of CMIP5 centennial simulations, *Geosci. Model Dev*., 4, 543-570, doi:10.5194/gmd-4-543-2011.

Kim, D., A.H. Sobel, A.D. Del Genio, Y. Chen, S. Camargo, M.-S. Yao, M. Kelley and L. Nazarenko, 2011: The tropical subseasonal variability simulated in the NASA GISS general circulation model. *Journal of Climate*, submitted.

Klein, S. A., and C. Jakob (1999), Validation and sensitivities of frontal clouds simulated by the ECMWF model, *Mon. Weather Rev.*, 127(10), 2514-2531.

L'Ecuyer, T.S., and J.H. Jiang, Touring the atmosphere aboard the A-Train, *Physics Today*, 63, 7, 36-41, 2010.

Li, J-L., et al. (2005), Comparisons of EOS MLS Cloud Ice Measurements with ECMWF analyses and GCM Simulations: Initial Results, *Geophys. Res. Lett.* 32, L18710, doi:10.1029/2005GL023788.

Li, J-L., et al. (2008), Comparisons of satellite liquid water estimates with ECMWF and GMAO analyses, 20th century IPCC AR4 climate simulations, and GCM simulations, *Geosphys. Res. Lett.*, 35, L19710, doi:10.1029/2008GL035427.

Livesey, N. J., et al. (2007), EOS MLS version 2.2 Level 2 data quality and description document, *Tech. Rep. JPL D-33509*, Jet Propul. Lab, Pasadena, Calif.

Martin, G. M., Milton, S. F., Senior, C. A., Brooks, M. E., Ineson, S., Reichler, T., and Kim, J. (2010): Analysis and Reduction of Systematic Errors through a Seamless Approach to Modelling Weather and Climate, *J. Climate*, 23, 5933–5957, doi:10.1175/2010 JCLI3541.1

Martin, G.M., M.A. Ringer, V.D. Pope, A. Jones, C. Dearden and T.J. Hinton (2006): The physical properties of the atmosphere in the new Hadley Centre Global Environmental Model, HadGEM1. Part 1: Model description and global climatology. *J. Climate*, 19, 1274-1301

Ming, Y., V. Ramaswamy, L.J. Donner, and V.T.J. Phillips (2006) A robust parameterization of cloud droplet activation. *J. Atmos. Sci., 63,* 1348-1356

Neale, R. B. et al (2010), Description of the NCAR Community Atmosphere Model (CAM5.0), *NCAR Technical Note, NCAR/TN-486+STR*. It is available at: (http://www.cesm.ucar.edu/models/cesm1.0/cam/docs/description/cam5_desc.pdf).

Olsen, E.T., S. Granger, E. Manning, J. Blaisdell (2007), AIRS/AMSU/HSB Version 5 Level 3 Quick Start, http://airs.jpl.nasa.gov/AskAirs.

Randall, D. A., et al. (2007), Climate models and their evaluations, in *Climate Change 2007: The Physical Sciences Basis, Contribution of Working Group I to the Fouth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., chapter 8, 589-662, Cambridge Univ. Press, U.K.

Randall, D. A., and S. Tjemkes (1991), Clouds, the Earth's radiation budget, and the hydrologic cycle, *Global and Planetary Change,* Volume 4, Issues 1-3, 3-9.

Read, W.G., Z. Shippony, M.J. Schwartz, N.J. Livesey, and W.V. Snyder (2006), The clear-sky unpolarized forward model for the EOS Microwave Limb Sounder (MLS), *IEEE Trans. Geosci. Remote Sensing* 44, no. 5, 1367-1379.

Read, W.G., et al. (2007) Aura Microwave Limb Sounder upper tropospheric and lower stratospheric H2O and relative humidity with respect to ice validation, *J. Geophys. Res.* 112, D24S35, doi:10.1029/2007JD008752.

Reichler, T. and Kim, J. (2008): How well do coupled models simulate today's climate?, *B. Am. Meteor. Soc*., 89 (3), 303–311, doi:10.1175/BAMS-89-3-303.

Rotstayn, L. D., 1997: A physically based scheme for the treatment of stratiform clouds and precipitation in large-scale models. I: Description and evaluation of the microphysical processes. *Q. J. R. Meteorol. Soc*., 123, 1227–1282.

Rotstayn, L. D., 1999: Climate sensitivity of the CSIRO GCM: Effect of cloud modeling assumptions. *J. Clim*., 12, 334–356.

Rotstayn LD, Collier MA, Dix MR, Feng Y, Gordon HB, O'Farrell SP, Smith IN, Syktus J. (2010), Improved simulation of Australian climate and ENSO-related climate variability in a GCM with an interactive aerosol treatment. *International Journal of Climatology*, 30:1067–1088. doi:10.1002/joc.1952.

von Salzen, K., J. F. Scinocca, N. A. McFarlane, J. Li, J. N. S. Cole, D. Plummer, M. C. Reader, M. Lazare, L. Solheim, Responses of Clouds and Precipitation to Short-Term Climate Variability in the Canadian Fourth Generation Atmospheric Global Circulation Model (CanAM4) (2012), in preparation.

Sakamato, T.T., and Coauthors, (2011) MIROC4h - a new high resolution atmosphere-ocean coupled general circulation model. *J. Meteor. Soc. Japan*, 89A, in press.

Soden, B. J., and I. M. Held (2006), An assessment of climate feedbacks in coupled ocean-atmosphere models, *J. Clim.*, 19, 3354-3360.

Su, H., D.E. Waliser, J.H. Jiang, J-L. Li, W.G. Read, J.W. Waters, and A.M. Tompkins, (2006), Relationships of upper tropospheric water vapor, clouds and SST: MLS observations, ECMWF analyses and GCM simulations, *Geophys. Res. Lett.* 33, L22802, doi:10.1029/2006GL027582.

Taylor, K.E.: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, 106, 7183-7192, 2001

Tiedtke, M. (1989), A comprehensive mass flux scheme for cumulus parametrization in large scale models, *Mon. Weather Rev.*, 117(8), 1779-1800.

Tiedtke, M. (1993), Representation of clouds in large-scale models, *Mon. Weather Rev.*, 121, 3040-3061.

Voldoire, A. et al. (2011), The CNRM-CM5.1 global climate model: Description and basic evaluation, *CNRM-CM technical doc*, available at http://www.cnrm.meteo.fr/cmip5/.

Volodin E.M., Diansky N.A., 2004. El-Nino reproduction in coupled general circulation model of atmosphere and ocean. *Russian meteorology and hydrology*, N 12, p.5-14.

Waliser, D.E., et al. (2009), Cloud ice: A climate model challenge with signs and expectations of progress, *J. Geophys. Res.* 114, D00A21, doi:10.1029/2008JD010015.

Watanabe, M., and Coauthors (2010), Improved climate simulation by MIROC5: Mean states, variability, and climate sensitivity. *J. Climate*, 23, 6312-6335.

Webb, M., C. Senior, S. Bony, and J. J. Morcrette (2011), Combining ERBE and ISCCP data to assess clouds in the Hadley Centre, ECMWF and LMD atmospheric climate models, *Clim. Dyn.*, 17(12), 905-922, doi:10.1007/s003820100157.

Woods, C. D. Waliser, J-L. Li, R. Austin, G. Stephens, and D. Vane (2008), Evaluating CloudSat ice water retrievals using a cloud-resolving model: Sensitivities to frozen particle properties, *J. Geosphys. Res.*, 113, D00A11, doi: 10.1029/2008JD009941.

Wu, D.L., J.H. Jiang, and C.P. Davis, EOS MLS cloud ice measurements and cloudy-sky radiative transfer model, *IEEE Trans. Geosci. Remote Sensing* 44, no. 5, 1156-1165.

Wu, D.L., J.H. Jiang, W.G. Read, R.T. Austin, C.P. Davis, A. Lambert, G.L. Stephens, D.G. Vane, and J.W. Waters (2008), Validation of the Aura MLS Cloud Ice Water Content (IWC) Measurements, *J. Geophys. Res.* 113, doi:10.1029/2007JD008931.

Wu, T. and Co-authors, 2011: The 20th century global carbon cycle from the Beijing Climate Center Climate System Model (BCC_CSM), submitted to Climate Dynamics.

Wu, T., R. Yu, F. Zhang, Z. Wang, M. Dong, L. Wang, X. Jin, D. Chen, L. Li (2010), The Beijing Climate Center for Atmospheric General Circulation Model (BCC-AGCM2.0.1): Description and its performance for the present-day climate, *Clim. Dyn*., 34: 123-147.

Wu, T., R. Yu, and F. Zhang, 2008: A modified dynamic framework for atmospheric spectral model and its application. J. Atmos. Sci., 65: 2235-2253.

Yukimoto, S., H. Yoshimura, M. Hosaka, T. Sakami, H. Tsujino, M. Hirabara, T. Y. Tanaka, M. Deushi, A. Obata, H. Nakano, Y. Adachi, E. Shindo, S. Yabu, T. Ose, and A. Kitoh, (2011a), Meteorological Research Institute Earth System Model Version 1 (MRI-ESM1) - Model Description -. *Technical Report of the Meteorological Research Institute*, 64, 83pp (available from http://www.mri-jma.go.jp/Publish/Technical/index_en.html).

Yukimoto, S. et al. (2011b), A New Global Climate Model of the Meteorological Research Institute: MRI-CGCM3 -Model Description and Basic Performance-, *J. Meteor. Soc. Japan*, in press.

Zhang, G. J., and N. A. McFarlane (1995), Sensitivity of climate simulations to the parameterization of cumulus convection in the Canadian Climate Centre General Circulation Model, *Atmos. Ocean*, 33, 407-446.

Table 1: AR5 and AR4 models used in this study

Modeling Center		AR4 Model ('20c3m' run)	AR5 Model ('historical' or AMIP run)	Type (AR5)	Resolution (AR5)	Key Reference (AR5)
Beijing Climate Center, China	BCC	-	csm1.1	AOGCM	2.8125°×2.8125°, L26	Wu et al. [2010] Wu et al. [2011]
Bjerknes Centre for Climate Research, Norway / Norwegian Climate Center, Norway	[1]BCCR-NCC	bcm2	noresm	AOGCM	2.5°×1.8947°, L26	Seland et al. [2008] Kirkevag et al. [2008]
Canadian Centre for Climate Modeling and Analysis, Canada	CCCMA	cgcm3.1	am4, canesm2	AOGCM AOGCM	2.8125°×2.7673°, L35	Arora et al. [2011] Salzen et al. [2012]
Centre National de Recherches Météorologiques, France	CNRM	cm3	cm5	AOGCM	1.4°×1.4°, L31	Voldoire et al. [2011]
Commonwealth Scientific and Industrial Research Organization / Queensland Climate Change Centre of Excellence, Australia	[2]CSIRO-QCCCE	mk3	mk3.6	AOGCM	1.9°×1.9°, L18	Rotstayn et al. [2010]
Geophysical Fluid Dynamics Laboratory, USA	GFDL	cm2	am3, cm3	AGCM AOGCM	2.5°×2°, L23	Donner et al. [2011] GFDL-AMDT [2011]
Goddard Institute for Space Studies, USA	GISS	e-h, e-r	e2-h, e2-r	AGCM AOGCM	5°×5°, L29	Kim et al. [2011]
Institute for Numerical Mathematics, Russia	INM	cm3	cm4	AOGCM	5°×4°, L21	Diansky et al. 2002. Diansky & Volodin 2002. Volodin & Diansky [2004]
Institut Pierre Simon Laplace, France	IPSL	cm4	cm5a	AOGCM	3.75°×1.8947°, L39	Dufresne et al. [2011] Hourdin et al. [2006]
Model for Interdisciplinary Research On Climate --- developed at Atmos. Ocean Res. Ins. (AORI), U. Tokyo / Nat. Ins. Env. Std. / Japan Agency for Marine-Earth Sci. & Tech., Japan	MIROC	[3]miroc3.2-medres	miroc4h, miroc5	AOGCM	0.5625°×0.55691°, L56; 1.4°×1.4°, L40	Watanabe et al. [2010] Sakamato et al. [2011]
Meteorological Research Institute, Japan	MRI	-	cgcm3	AOGCM	1.125°×1.1121°, L35	Yukimoto et al. [2011a] Yukimoto et al. [2011b]
National Center for Atmospheric Research, USA	NCAR	ccsm3	[4]cam5-cesm1	AOGCM	1.25°×0.9424°, L30	Eaton [2011] Neale et al. [2010]
UK Met Office, Hadley Climate Center, UK	UKMO	hadgem1	hadgem2-es, hadgem2-a hadgem2-cc	AOGCM AGCM AOGCM	1.875°×1.25°, L38	HDT [2011] Collins et al. [2011] Jones et al. [2011]

Note: 1, 2, 3, 4: For simplicity, acronyms "BCCR", "CSIRO", "miroc3.2", and "cam5" will be used in the text for model descriptions.

Table 2: Model outputs used in this study

AR5 Model Variable	**Acronym** (unit)	**Note**
Ice Water Path (2D)	clivi (kg/m^2)	Mass of ice water in the column divided by area of column
Condensed Water Path (2D)	clwvi (kg/m^2)	Mass of condensed (liquid+ice) water in column divided by area of column
Mass fraction of cloud ice water (3D)	cli (kg/kg)	Mass fraction of cloud ice in atmospheric layer
Mass fraction of cloud liquid water (3D)	clw (kg/kg)	Mass fraction of cloud liquid water in atmospheric layer
Water Vapor Path (2D)	prw (kg/m^2)	Atmospheric water vapor content vertically integrated through the column
Specific humidity (3D)	hus (kg/kg)	Mass fraction atmospheric water vapor in atmospheric layer

Table 3: A-Train data products used in this study

Data source	**Data product**	**Acronym** (units)	**Estimated uncertainty**
Aqua AIRS	Water Vapor Mixing Ratio	H_2O (g/kg)	25-30%
Aqua AMSR-E	Water Vapor Path	WVP (kg/m^2)	20%
Aqua MODIS	Ice Water Path Liquid Water Path	IWP (g/m^2) LWP (g/m^2)	Factor of 2
Aura MLS	Water Vapor Mixing Ratio Ice Water Content	H_2O (ppmv) IWC (mg/m^3)	≤ 20% Factor of 2
CALIPSO	Ice Water Content	IWC (mg/m^3)	Factor of 2
CloudSat	Ice Water Content Liquid Water Content	IWC (mg/m^3) LWC (mg/m^3)	Factor of 2

Table 4a: Spatial means $\overline{IWC}_{mdl}$ / $\overline{LWC}_{mdl}$ and spatial mean scores $G_m^{IWC\,/\,LWC}$ for IWC and LWC. Observed means and their uncertainty ranges are immediately below the labels in the top row.

AR5 Model	**100 hPa** (MLS) **0.0438** (0.0219-0.0875) **mg/m³**		**215 hPa** (MLS) **2.39** (1.20-4.78) **mg/m³**		**600 hPa** (CloudSat) **2.77** (1.27-5.55) **mg/m³**		**900 hPa** (CloudSat) **24.4** (3.06-48.8) **mg/m³**	
	$\overline{IWC}_{mdl}$	G_m^{IWC}	$\overline{IWC}_{mdl}$	G_m^{IWC}	$\overline{LWC}_{mdl}$	G_m^{LWC}	$\overline{LWC}_{mdl}$	G_m^{LWC}
BCC csm1	0.00851	0.21	0.460	0.21	9.16	0.43	18.4	0.90
BCCR noresm	0.0328	0.86	0.974	0.57	9.09	0.43	15.1	0.83
CCCMA am4	0.00505	0.0	2.39	1.0	5.52	0.67	27.9	0.95
CCCMA canesm2	0.00523	0.0	2.44	0.99	6.05	0.63	30.8	0.92
CNRM cm5	0.00338	0.0	1.09	0.62	8.79	0.45	18.0	0.89
CSIRO mk3.6	0.0139	0.45	1.03	0.60	2.79	1.0	23.5	0.99
GFDL am3	1.01	0.0	6.98	0.48	5.63	0.66	15.5	0.84
GFDL cm3	0.646	0.0	6.75	0.50	5.72	0.65	16.3	0.85
GISS e2-h	0.0234	0.70	22.9	0.0	4.69	0.75	17.9	0.89
GISS e2-r	0.0354	0.90	23.8	0.0	4.57	0.76	15.7	0.84
INM cm4	0.00393	0.0	0.0729	0.0	1.75	0.78	4.53	0.39
IPSL cm5a	0.0133	0.43	2.51	0.98	6.26	0.61	11.8	0.74
MIROC miroc4h	0.0918	0.64	3.04	0.88	8.91	0.44	48.2	0.75
MIROC miroc5	0.00347	0.0	1.20	0.67	8.05	0.49	42.7	0.80
MRI cgcm3	0.00868	0.22	1.86	0.88	10.9	0.34	11.9	0.74
NCAR cam5	0.00356	0.0	1.37	0.73	0.940	0.48	12.6	0.76
UKMO hadgem2-a	0.00607	0.05	1.47	0.77	2.63	0.97	17.8	0.89
UKMO hadgem2-cc	0.00330	0.0	1.20	0.67	2.98	0.97	18.5	0.90
UKMO hadgem2-es	0.00389	0.0	1.28	0.70	2.83	0.99	17.9	0.89

Table 4b: Model spatial means $\overline{H_2O}_{mdl}$ and spatial mean scores $G_m^{H_2O}$ for H_2O. Observed means and their uncertainty ranges are immediately below the labels in the top row.

AR5 Model	**100 hPa** (MLS) **0.259** (±0.0259)10^{-2} **g/kg**		**215 hPa** (MLS) **0.466** (±0.0932)10^{-1} **g/kg**		**600 hPa** (AIRS) **2.58** (±0.646) **g/kg**		**900 hPa** (AIRS) **11.5** (±2.88) **g/kg**	
	$\overline{H_2O}_{mdl}$	$G_m^{H_2O}$	$\overline{H_2O}_{mdl}$	$G_m^{H_2O}$	$\overline{H_2O}_{mdl}$	$G_m^{H_2O}$	$\overline{H_2O}_{mdl}$	$G_m^{H_2O}$
BCC csm1	0.217	0.47	0.462	0.99	2.46	0.94	10.3	0.85
BCCR noresm	0.261	0.97	0.623	0.44	2.81	0.88	10.6	0.89
CCCMA am4	0.241	0.78	0.754	0.0	2.53	0.98	10.5	0.88
CCCMA canesm2	0.253	0.92	0.791	0.0	2.56	0.99	10.5	0.89
CNRM cm5	0.174	0.0	0.430	0.87	2.45	0.93	10.7	0.90
CSIRO mk3.6	0.360	0.0	0.868	0.0	2.87	0.85	10.9	0.93
GFDL am3	0.259	1.0	0.871	0.0	2.96	0.81	11.1	0.95
GFDL cm3	0.247	0.84	0.740	0.021	2.70	0.94	10.7	0.90
GISS e2-h	0.348	0.0	0.702	0.16	2.35	0.88	11.6	0.99
GISS e2-r	0.371	0.0	0.820	0.0	2.50	0.96	11.9	0.96
INM cm4	0.378	0.0	0.466	1.0	3.30	0.63	10.4	0.87
IPSL cm5a	0.168	0.0	0.654	0.33	2.67	0.95	9.35	0.75
MIROC miroc4h	0.206	0.31	0.709	0.13	2.51	0.96	10.1	0.84
MIROC miroc5	0.00181	0.0	0.0561	0.66	2.64	0.97	10.8	0.92
MRI cgcm3	0.395	0.0	0.747	0.0	3.14	0.71	11.2	0.96
NCAR cam5	0.231	0.65	0.593	0.55	3.11	0.73	12.0	0.95
UKMO hadgem2-a	0.304	0.42	0.510	0.85	2.63	0.98	10.9	0.92
UKMO hadgem2-cc	0.252	0.91	0.407	0.79	2.35	0.88	10.2	0.85
UKMO hadgem2-es	0.292	0.57	0.442	0.92	2.42	0.92	10.4	0.87

Table 5a: Model spatial standard deviations $\sigma_{mdl}^{IWC/LWC}$ (normalized to the observed spatial standard deviation), and spatial variance scores $G_{\mathrm{v}}^{IWC/LWC}$, for IWC and LWC.

AR5 Model	100 hPa		215 hPa		600 hPa		900 hPa	
	σ_{mdl}^{IWC}	G_{v}^{IWC}	σ_{mdl}^{IWC}	G_{v}^{IWC}	σ_{mdl}^{LWC}	G_{v}^{LWC}	σ_{mdl}^{LWC}	G_{v}^{LWC}
BCC csm1	0.137	0.043	0.0949	0.0	2.88	0.49	0.384	0.66
BCCR noresm	0.744	0.86	0.244	0.32	3.09	0.46	0.567	0.80
CCCMA am4	0.117	0.0	0.869	0.93	3.42	0.41	0.701	0.87
CCCMA canesm2	0.137	0.042	0.911	0.96	3.73	0.37	0.827	0.93
CNRM cm5	0.0989	0.0	0.418	0.58	3.18	0.44	0.388	0.66
CSIRO mk3.6	0.186	0.19	0.410	0.57	1.08	0.96	0.842	0.94
GFDL am3	27.7	0.0	2.893	0.49	2.13	0.64	0.382	0.65
GFDL cm3	17.1	0.0	2.570	0.55	1.93	0.68	0.320	0.59
GISS e2-h	1.77	0.72	10.1	0.0	3.94	0.34	0.422	0.69
GISS e2-r	2.86	0.50	10.7	0.0	3.56	0.39	0.488	0.74
INM cm4	0.0666	0.0	0.0216	0.0	0.578	0.74	0.0767	0.074
IPSL cm5a	0.333	0.47	0.807	0.90	2.88	0.49	0.478	0.73
MIROC miroc4h	1.83	0.71	1.12	0.95	3.86	0.35	0.920	0.97
MIROC miroc5	0.0592	0.0	0.433	0.60	3.56	0.39	0.666	0.85
MRI cgcm3	0.222	0.28	0.674	0.81	3.84	0.35	0.221	0.46
NCAR cam5	0.0929	0.0	0.492	0.66	0.451	0.62	0.668	0.86
UKMO hadgem2-a	0.173	0.16	0.499	0.67	0.930	0.97	0.562	0.79
UKMO hadgem2-cc	0.0936	0.0	0.407	0.57	0.996	1.0	0.449	0.71
UKMO hadgem2-es	0.116	0.0	0.437	0.60	0.983	0.99	0.462	0.72

Table 5b: Model spatial standard deviations $\sigma_{mdl}^{H_2O}$ (normalized to the observed spatial standard deviation), and spatial variance scores $G_{\mathrm{v}}^{H_2O}$, for H_2O.

AR5 Model	100 hPa		215 hPa		600 hPa		900 hPa	
	$\sigma_{mdl}^{H_2O}$	$G_{\mathrm{v}}^{H_2O}$	$\sigma_{mdl}^{H_2O}$	$G_{\mathrm{v}}^{H_2O}$	$\sigma_{mdl}^{H_2O}$	$G_{\mathrm{v}}^{H_2O}$	$\sigma_{mdl}^{H_2O}$	$G_{\mathrm{v}}^{H_2O}$
BCC csm1	1.53	0.0	0.476	0.13	0.671	0.56	0.846	0.80
BCCR noresm	1.06	0.81	0.830	0.72	1.09	0.88	0.880	0.84
CCCMA am4	2.55	0.0	1.36	0.40	0.872	0.83	1.01	0.98
CCCMA canesm2	2.68	0.0	1.46	0.24	0.911	0.88	1.06	0.92
CNRM cm5	0.887	0.62	0.544	0.24	0.881	0.84	0.819	0.76
CSIRO mk3.6	3.17	0.0	1.43	0.29	1.12	0.85	1.03	0.96
GFDL am3	2.18	0.0	1.54	0.10	1.26	0.65	1.00	1.0
GFDL cm3	2.20	0.0	1.16	0.73	1.02	0.98	0.877	0.84
GISS e2-h	1.34	0.0	0.993	0.99	0.754	0.67	0.951	0.94
GISS e2-r	1.63	0.0	1.28	0.53	0.881	0.84	1.10	0.87
INM cm4	5.43	0.0	0.754	0.59	1.21	0.71	0.891	0.86
IPSL cm5a	0.687	0.0	0.902	0.84	1.07	0.91	0.934	0.91
MIROC miroc4h	3.42	0.0	1.14	0.77	1.10	0.87	1.03	0.97
MIROC miroc5	2.45	0.0	0.706	0.51	1.04	0.95	0.873	0.83
MRI cgcm3	1.58	0.0	1.05	0.92	1.02	0.98	0.904	0.87
NCAR cam5	1.38	0.0	0.788	0.65	1.18	0.75	0.868	0.83
UKMO hadgem2-a	1.27	0.11	0.816	0.69	1.01	0.99	0.871	0.83
UKMO hadgem2-cc	1.08	0.73	0.644	0.41	0.897	0.86	0.802	0.74
UKMO hadgem2-es	1.27	0.11	0.716	0.53	0.939	0.92	0.828	0.77

Table 6a: Model-observation spatial correlation coefficients $C_{mdl,obs}^{IWC/LWC}$, and model spatial correlation scores $G_c^{IWC/LWC}$, for IWC/LWC.

AR5 Model	100 hPa		215 hPa		600 hPa		900 hPa	
	$C_{mdl,obs}^{IWC}$	G_c^{IWC}	$C_{mdl,obs}^{IWC}$	G_c^{IWC}	$C_{mdl,obs}^{LWC}$	G_c^{LWC}	$C_{mdl,obs}^{LWC}$	G_c^{LWC}
BCC csm1	0.706	0.71	0.812	0.81	0.613	0.61	0.229	0.23
BCCR noresm	0.592	0.59	0.814	0.81	0.645	0.64	0.434	0.43
CCCMA am4	0.831	0.83	0.813	0.81	0.367	0.37	0.377	0.38
CCCMA canesm2	0.728	0.73	0.784	0.78	0.371	0.37	0.336	0.34
CNRM cm5	0.613	0.61	0.830	0.83	0.661	0.66	0.143	0.14
CSIRO mk3.6	0.664	0.66	0.818	0.82	0.601	0.60	0.751	0.75
GFDL am3	0.818	0.82	0.894	0.89	0.812	0.81	0.729	0.73
GFDL cm3	0.746	0.75	0.794	0.79	0.662	0.66	0.639	0.64
GISS e2-h	0.258	0.26	0.642	0.64	−0.0294	0.00	0.479	0.48
GISS e2-r	0.241	0.24	0.677	0.68	−0.0364	0.00	0.523	0.52
INM cm4	0.581	0.58	0.492	0.49	0.507	0.51	0.227	0.23
IPSL cm5a	0.629	0.63	0.779	0.78	0.687	0.69	0.497	0.50
MIROC miroc4h	0.849	0.85	0.834	0.83	0.658	0.66	0.471	0.47
MIROC miroc5	0.694	0.69	0.865	0.87	0.759	0.76	0.384	0.38
MRI cgcm3	0.632	0.63	0.788	0.79	0.697	0.70	0.205	0.21
NCAR cam5	0.842	0.84	0.857	0.86	0.576	0.58	0.488	0.49
UKMO hadgem2-a	0.677	0.68	0.831	0.83	0.620	0.62	0.636	0.64
UKMO hadgem2-cc	0.732	0.73	0.893	0.89	0.736	0.74	0.477	0.48
UKMO hadgem2-es	0.717	0.72	0.896	0.90	0.716	0.72	0.550	0.55

Table 6b: Model-observation spatial correlation coefficients $C_{mdl,obs}^{H_2O}$, and model spatial correlation scores $G_c^{H_2O}$, for H_2O.

AR5 Model	100 hPa		215 hPa		600 hPa		900 hPa	
	$C_{mdl,obs}^{H_2O}$	$G_c^{H_2O}$	$C_{mdl,obs}^{H_2O}$	$G_c^{H_2O}$	$C_{mdl,obs}^{H_2O}$	$G_c^{H_2O}$	$C_{mdl,obs}^{H_2O}$	$G_c^{H_2O}$
BCC csm1	0.805	0.80	0.845	0.85	0.882	0.88	0.929	0.93
BCCR noresm	0.0383	0.04	0.867	0.87	0.878	0.88	0.924	0.92
CCCMA am4	−0.075	0.00	0.898	0.90	0.921	0.92	0.946	0.95
CCCMA canesm2	−0.159	0.00	0.881	0.88	0.916	0.92	0.950	0.95
CNRM cm5	0.807	0.81	0.889	0.89	0.931	0.93	0.945	0.95
CSIRO mk3.6	0.569	0.57	0.890	0.89	0.888	0.89	0.961	0.96
GFDL am3	0.842	0.84	0.941	0.94	0.975	0.98	0.964	0.96
GFDL cm3	0.797	0.80	0.864	0.86	0.889	0.89	0.921	0.92
GISS e2-h	−0.152	0.00	0.738	0.74	0.800	0.80	0.893	0.89
GISS e2-r	−0.221	0.00	0.764	0.76	0.853	0.85	0.931	0.93
INM cm4	0.556	0.56	0.839	0.84	0.911	0.91	0.920	0.92
IPSL cm5a	0.494	0.49	0.893	0.89	0.894	0.89	0.911	0.91
MIROC miroc4h	0.558	0.56	0.857	0.86	0.912	0.91	0.957	0.96
MIROC miroc5	0.724	0.72	0.915	0.91	0.952	0.95	0.968	0.97
MRI cgcm3	0.807	0.81	0.809	0.81	0.833	0.83	0.889	0.89
NCAR cam5	0.789	0.79	0.913	0.91	0.975	0.97	0.955	0.96
UKMO hadgem2-a	0.892	0.89	0.857	0.86	0.935	0.94	0.963	0.96
UKMO hadgem2-cc	0.868	0.87	0.906	0.91	0.936	0.94	0.935	0.94
UKMO hadgem2-es	0.899	0.90	0.915	0.92	0.949	0.95	0.941	0.94

7: Overall scores and ranks for the AR5 models at individual pressure levels

AR5 Model	100 hPa		215 hPa		600 hPa		900 hPa	
	score	rank	score	rank	score	rank	score	rank
BCC csm1	0.37	7	0.50	12	0.65	10	0.73	10
BCCR noresm	0.69	1	0.62	9	0.70	8	0.79	7
CCCMA am4	0.27	12	0.67	7	0.70	8	0.84	3
CCCMA canesm2	0.28	11	0.64	8	0.69	9	0.82	5
CNRM cm5	0.34	8	0.67	7	0.71	7	0.72	11
CSIRO mk3.6	0.31	10	0.53	11	0.86	3	0.92	1
GFDL am3	0.44	4	0.49	13	0.76	5	0.86	2
GFDL cm3	0.40	5	0.58	10	0.80	4	0.79	7
GISS e2-h	0.28	11	0.42	14	0.57	12	0.81	6
GISS e2-r	0.27	12	0.33	15	0.63	11	0.81	6
INM cm4	0.19	14	0.49	13	0.71	7	0.56	13
IPSL cm5a	0.34	8	0.79	1	0.76	5	0.76	9
MIROC miroc4h	0.51	3	0.74	4	0.70	8	0.83	4
MIROC miroc5	0.24	13	0.70	6	0.75	6	0.79	7
MRI cgcm3	0.32	9	0.70	6	0.65	10	0.69	12
NCAR cam5	0.38	6	0.73	5	0.69	9	0.81	6
UKMO hadgem2-a	0.38	6	0.78	2	0.91	1	0.84	3
UKMO hadgem2-cc	0.54	2	0.70	6	0.90	2	0.77	8
UKMO hadgem2-es	0.38	6	0.76	3	0.91	1	0.79	7

Table 8: Overall scores and ranks for the AR5 models.

AR5 Model	Overall Score	Rank
BCC csm1	0.56	11
BCCR noresm	0.70	3
CCCMA am4	0.62	8
CCCMA canesm2	0.61	9
CNRM cm5	0.61	9
CSIRO mk3.6	0.65	6
GFDL am3	0.64	7
GFDL cm3	0.64	7
GISS e2-h	0.52	12
GISS e2-r	0.51	13
INM cm4	0.49	14
IPSL cm5a	0.66	5
MIROC miroc4h	0.69	4
MIROC miroc5	0.62	8
MRI cgcm3	0.59	10
NCAR cam5	0.65	6
UKMO hadgem2-a	0.73	1
UKMO hadgem2-cc	0.73	1
UKMO hadgem2-es	0.71	2
“Multi-model mean”	(0.78)	n/a

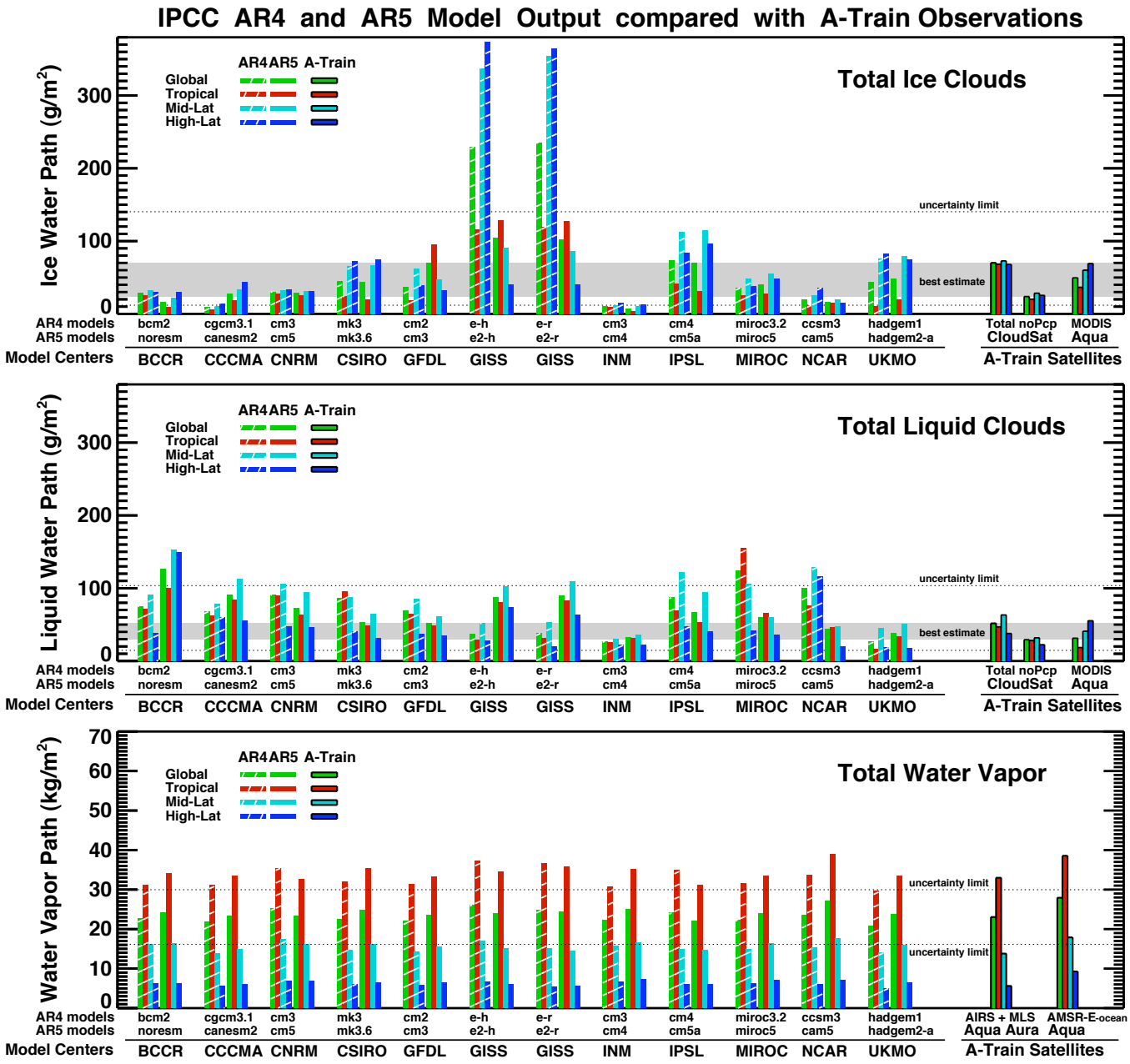

Figure 1. Multi-year global and zonal mean IWP, LWP, and WVP from AR4 and AR5 models, and from A-Train observations as described in the text.

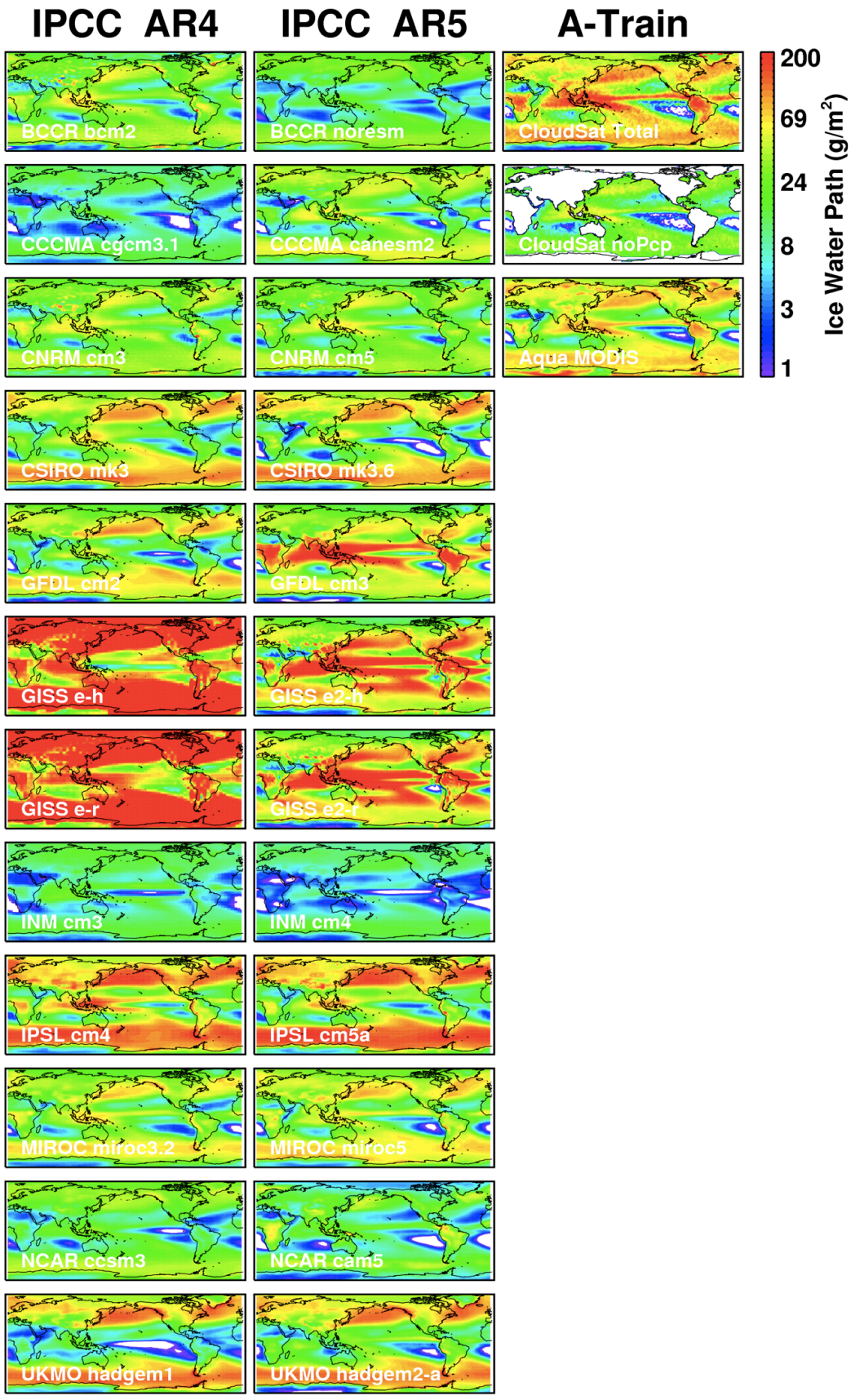

Figure 2a: Multi-year mean IWP from IPCC AR4 and AR5 models, and from A-Train observations.

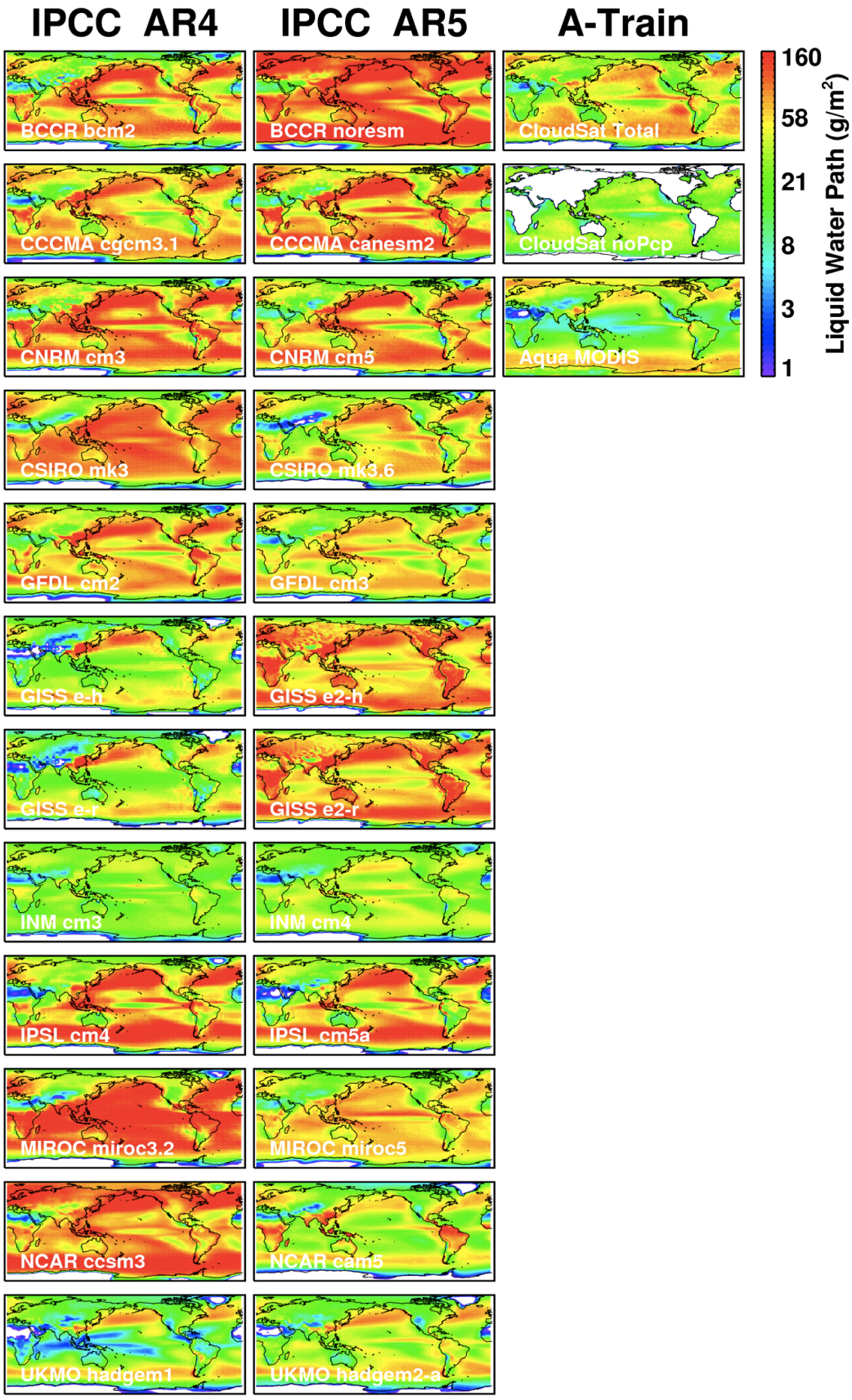

Figure 2b: Multi-year mean LWP from IPCC AR4 and AR5 models, and from A-Train observations.

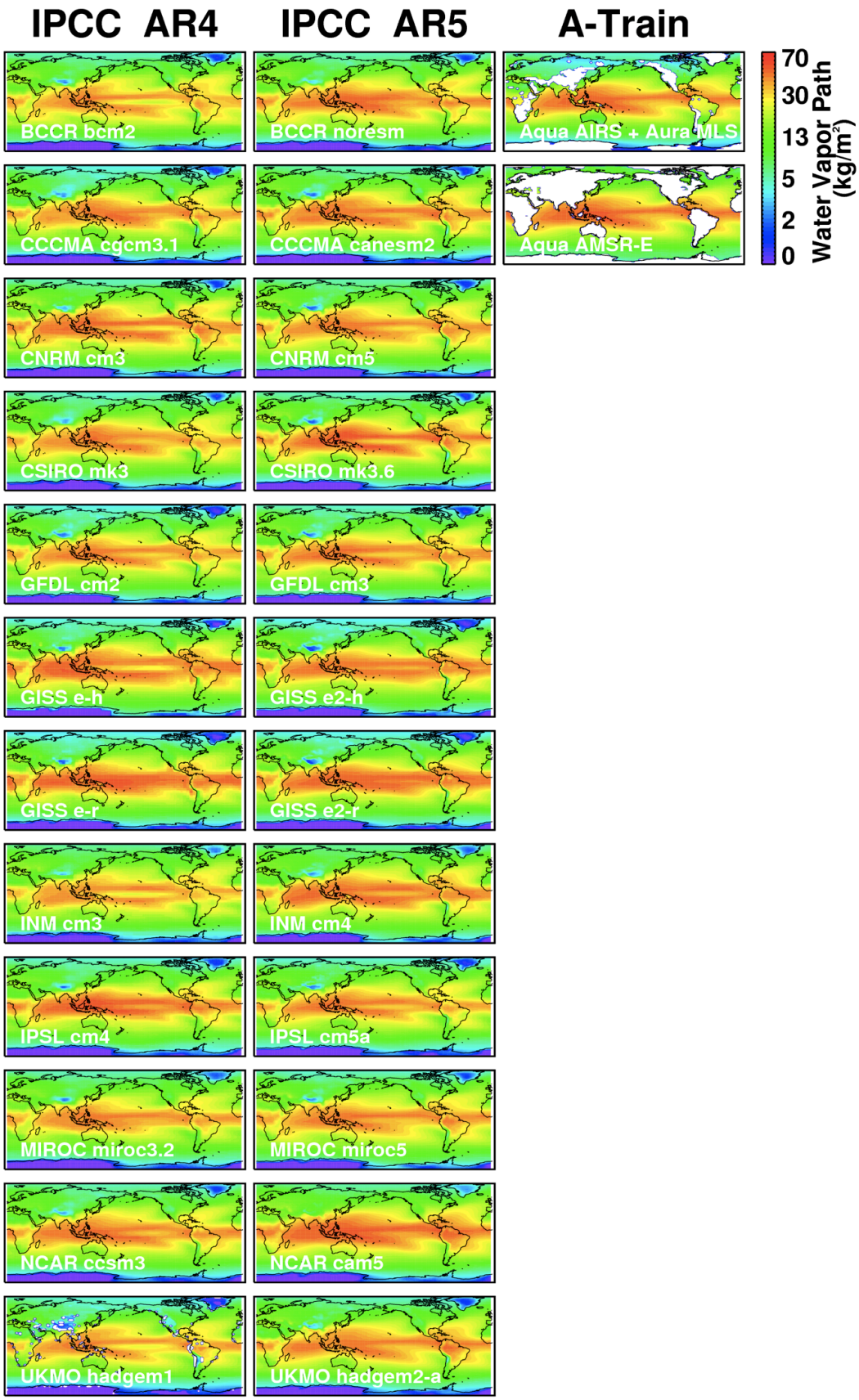

Figure 2c: Multi-year mean WVP from IPCC AR4 and AR5 models, and from A-Train observations.

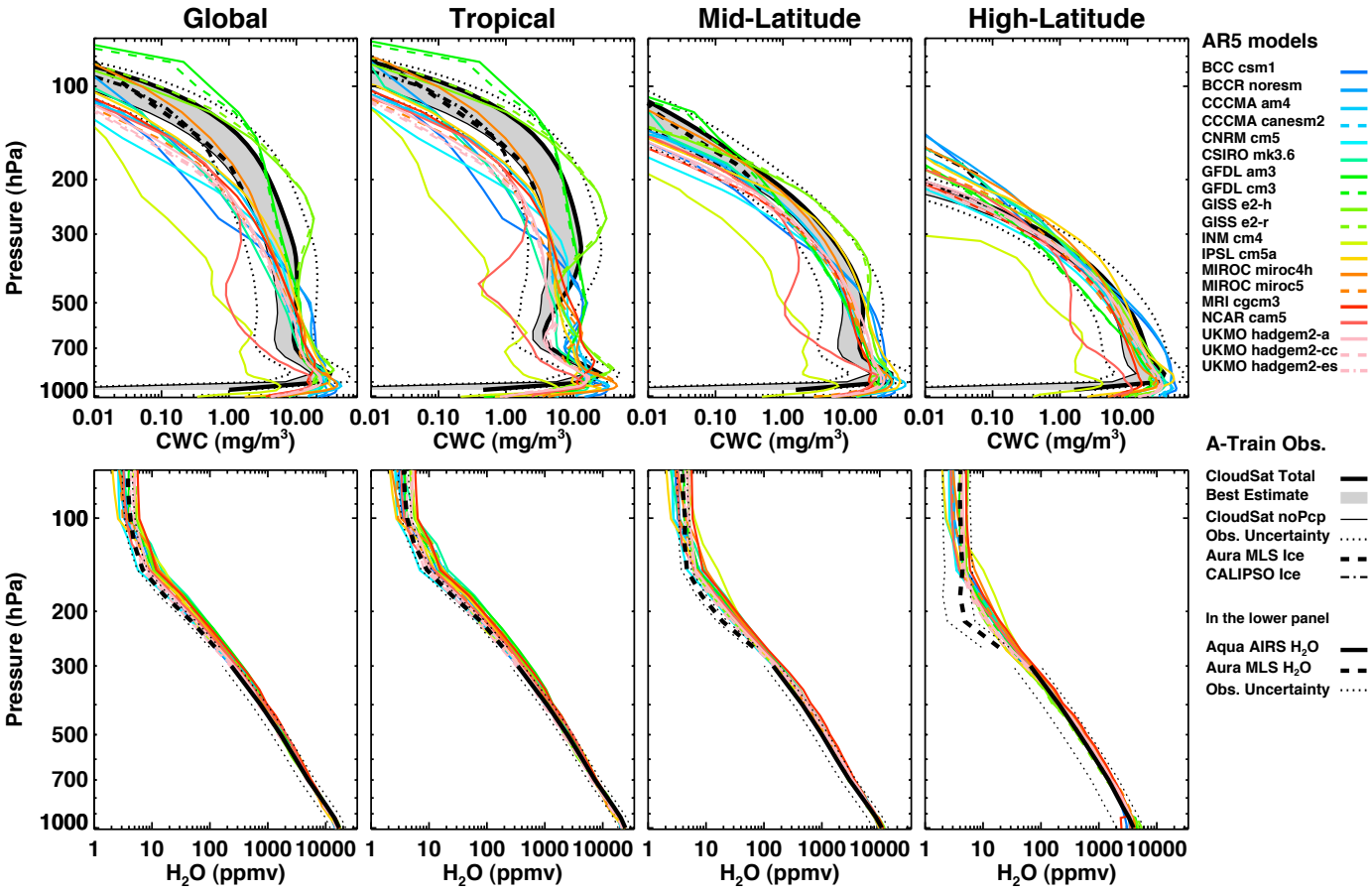

Figure 3a: Multi-year mean CWC and IWC (top panels) and H_2O (lower panels) vertical profiles from AR5 models and from A-Train observations. In the top panels, IWC is plotted for P ≤ 215 hPa.

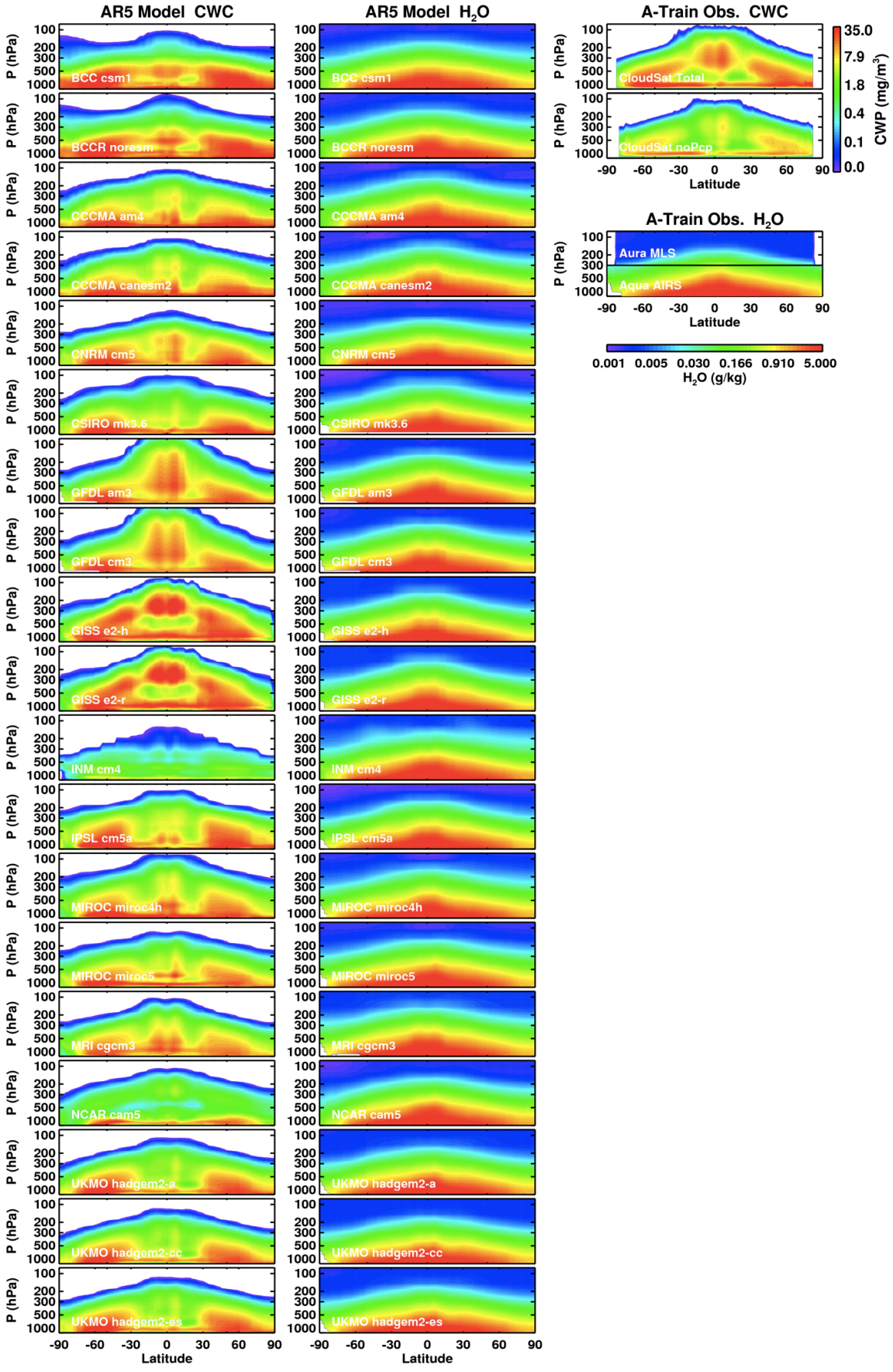

Figure 3b: Multi-year mean zonal profiles of CWC and H_2O from AR4/AR5 models and from A-Train observations. For Aura MLS observation, H_2O is plotted for P < 300 hPa, and for Aqua AIRS observation, H_2O is plotted for P ≥ 300 hPa

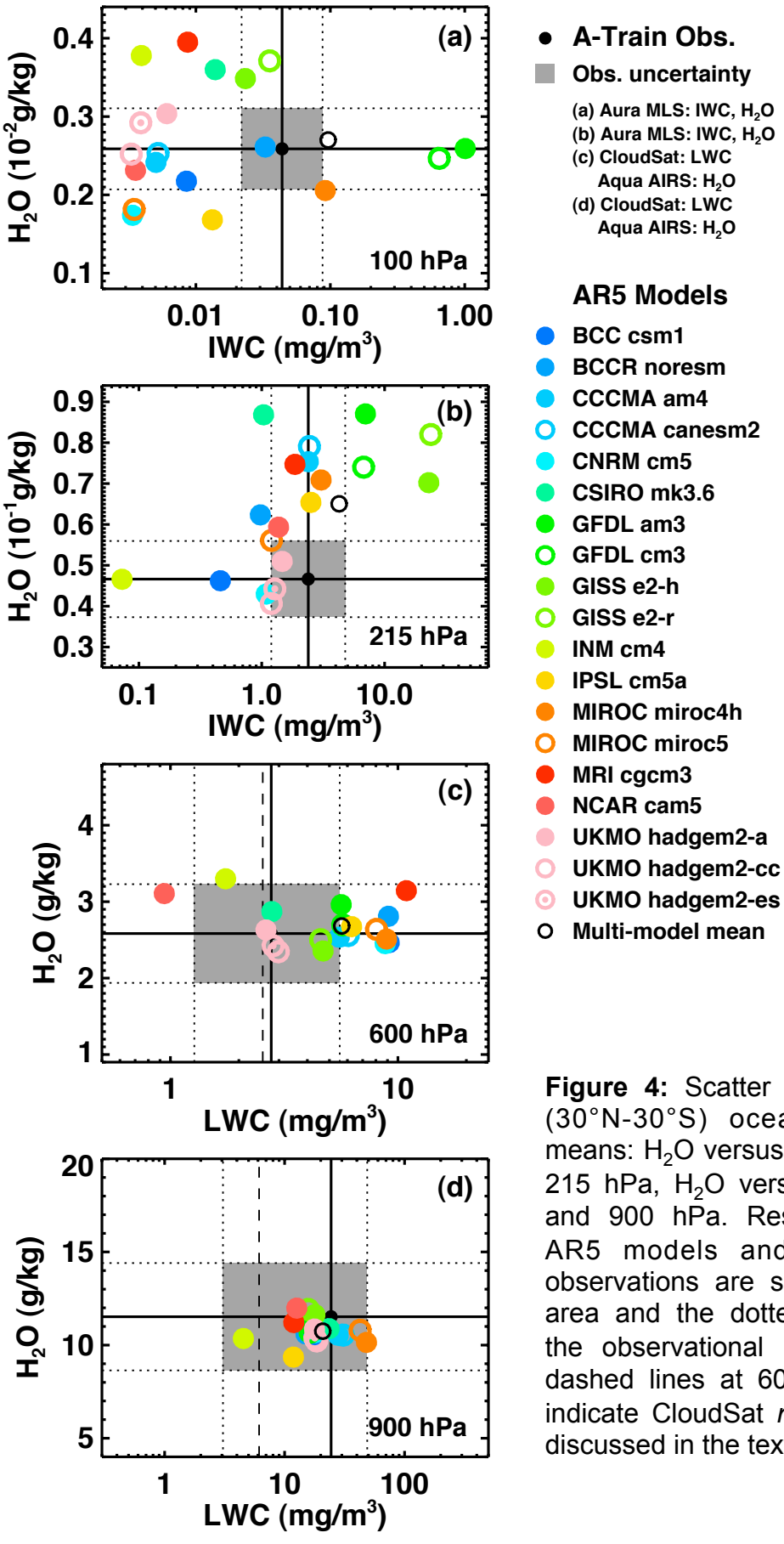

Figure 4: Scatter plots of tropical (30°N-30°S) oceanic multi-year means: H_2O versus IWC at 100 and 215 hPa, H_2O versus LWC at 600 and 900 hPa. Results from each AR5 models and from A-Train observations are shown. The grey area and the dotted lines indicate the observational uncertainty. The dashed lines at 600 and 900 hPa indicate CloudSat *noPcp* values as discussed in the text.

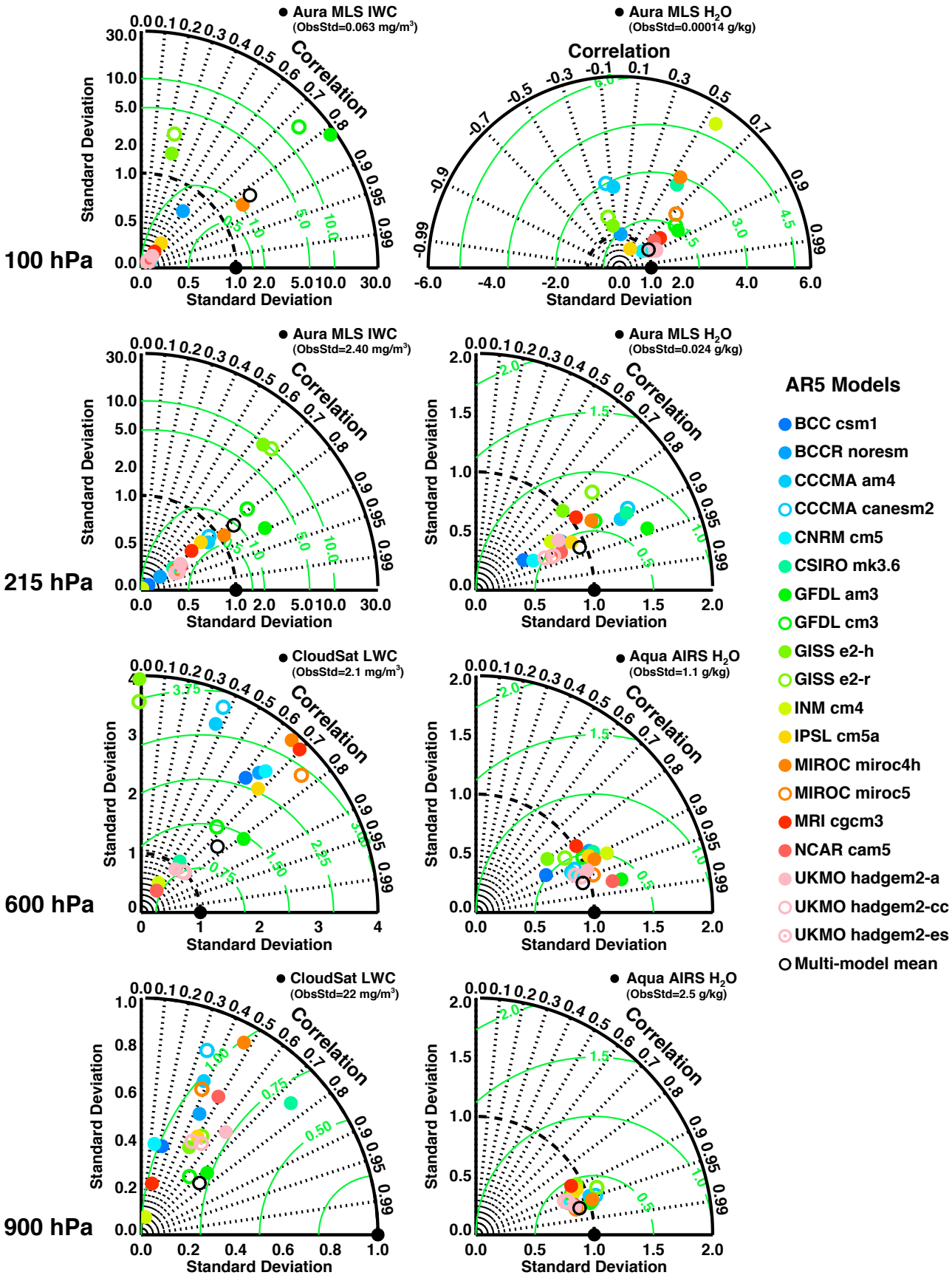

Figure 5: Taylor diagrams showing the tropical (30°N-30°S) oceanic multi-year mean performance of the AR5 models as compared to the A-Train observations. See text for more explanation.

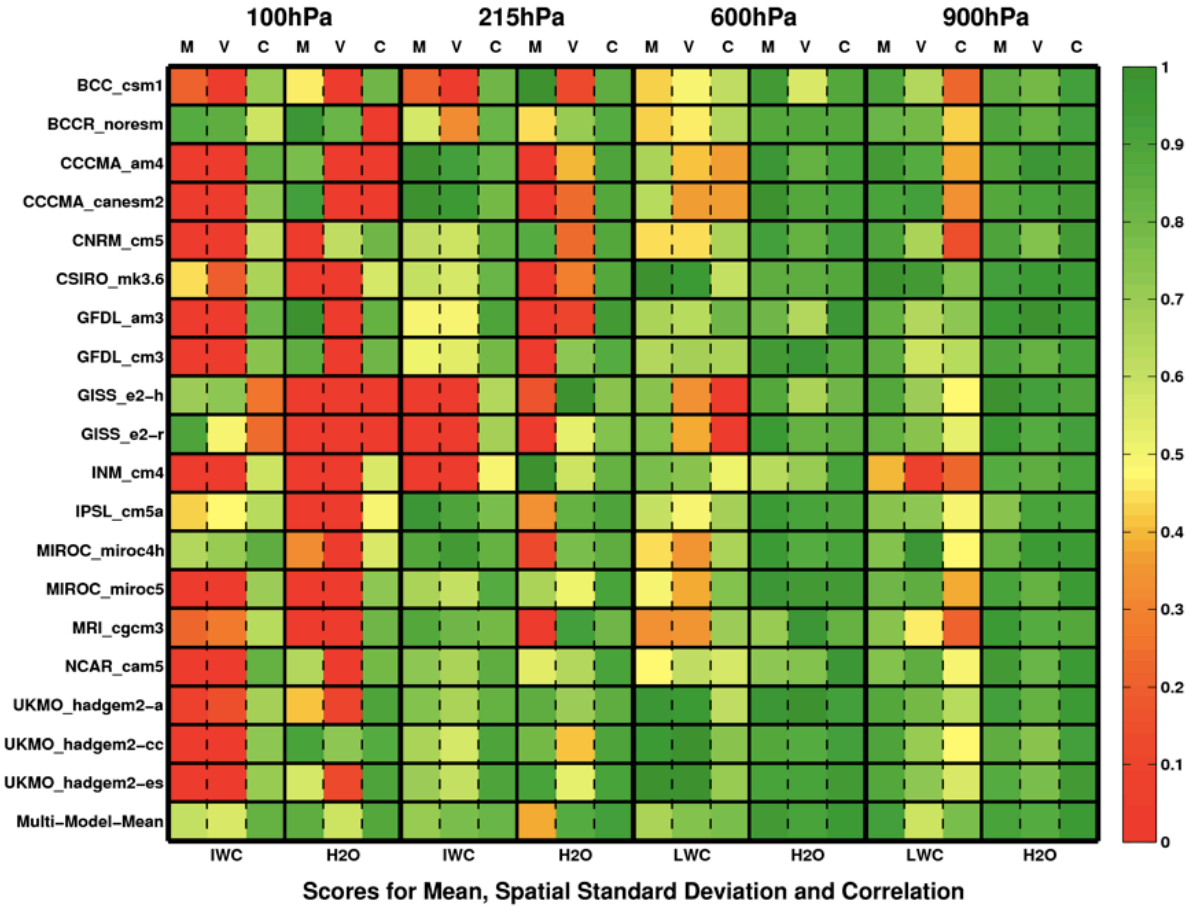

Figure 6: Color-coded summary of performance scores at 100, 215, 600, and 900 hPa. M: spatial mean performance scores G_m; V: spatial variance performance scores G_v; C: spatial correlation performance scores G_c.